

This is a repository copy of *Normative Requirements Operationalization with Large Language Models*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/219527/>

Version: Accepted Version

Proceedings Paper:

Feng, Nick, Marsso, Lina, Getir Yaman, Sinem et al. (9 more authors) (2024) Normative Requirements Operationalization with Large Language Models. In: Liebel, Grischa, Hadar, Irit and Spoletini, Paola, (eds.) Proceedings - 32nd IEEE International Requirements Engineering Conference, RE 2024. 32nd IEEE International Requirements Engineering Conference, RE 2024, 24-28 Jun 2024 Proceedings of the IEEE International Conference on Requirements Engineering . IEEE Computer Society , ISL , pp. 129-141.

<https://doi.org/10.1109/RE59067.2024.00022>

Reuse

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:

<https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

Normative Requirements Operationalization with Large Language Models

Abstract—Normative non-functional requirements specify constraints that a system must observe in order to avoid violations of social, legal, ethical, empathetic, and cultural norms. As these requirements are typically defined by non-technical system stakeholders with different expertise and priorities (ethicists, lawyers, social scientists, etc.), ensuring their well-formedness and consistency is very challenging. Recent research has tackled this challenge using a domain-specific language to specify normative requirements as rules whose consistency can then be analysed with formal methods. In this paper, we propose a complementary approach that uses Large Language Models to extract semantic relationships between abstract representations of system capabilities. These relations, which are often assumed implicitly by non-technical stakeholders (e.g., based on common sense or domain knowledge), are then used to enrich the automated reasoning techniques for eliciting and analyzing the consistency of normative requirements. We show the effectiveness of our approach to normative requirements elicitation and operationalization through a range of real-world case studies.

I. INTRODUCTION

Software systems are increasingly capable to perform daily tasks alongside humans, as important components of our infrastructure. Therefore, ensuring the *responsible* integration of these systems into our society is becoming crucial. This includes preventing them from violating social, legal, ethical, empathetic, and cultural (SLEEC) norms [45], [20], [21]. As such, it is important to elicit *normative non-functional requirements* (N-NFRs). Due to the nature of the N-NFRs, their elicitation involves non-technical stakeholders (e.g., ethicists, lawyers), each with different and potentially conflicting jargon, goals, priorities, and responsibilities. So, it is particularly important to support the professionals in ensuring that the elicited N-NFRs do not have well-formedness issues (WFIs) [20], [16], [17], [21] such as conflict and redundancy [21].

Previous work has proposed a rule-based SLEEC language to formalize N-NFRs [20]. Automated techniques, based on first-order logic [19] and a process algebra [22], [5], have been developed to check normative requirements specified in this language for WFIs [20], [21], [16], [17]. While effective, these techniques assume that all capabilities referenced in the rules are independent. More precisely, the non-terminal symbols used for describing events and measures in the rules are treated as uninterpreted constants that do not carry semantics. However, in practice, many of these capabilities are related, if we take into account common sense or domain knowledge, and those relations are often assumed by non-technical stakeholders. For example, ‘on’ and ‘off’ measures of a camera can be reasonably expected to be mutually exclusive. Ignoring these

semantic relations can affect the soundness and completeness of the WFI analysis. For example, consider the SLEEC rules:

```
 $r_1 = \text{when BatteryLow then MuteNotifications}$   
 $r_2 = \text{when BatteryLow then SendUserWarning}$ 
```

In the rule set $\{r_1, r_2\}$, r_1 is conflicting with r_2 . This is because r_1 and r_2 share the same triggering event `BatteryLow`, and the response of r_1 , `MuteNotifications`, is contradictory with the response of r_2 , `SendUserWarning`. The contradiction follows from the fact that, if the events `MuteNotifications` and `SendUserWarning` occur simultaneously, that is, in the same time unit, they cannot both have their intended effect. When all goes well (that is, there is no failure), either a warning is delivered, or the agent is muted, but not both. However, existing automated reasoning techniques [20], [16], [17] fail to detect this conflict as they do not capture the semantic relation between the two events used in the responses. Therefore, there is a need to define and capture semantic relations of terms used in N-NFRs according to common sense and domain knowledge, and to use these relations in the N-NFR analysis.

Traditional knowledge extraction approaches may be considered as a means for addressing this need. However, they use as input (structured or unstructured) sources without information specific to applications, while the semantic relations for capabilities used in SLEEC requirements come from common sense in the context of domain-specific environments. Therefore, these knowledge extraction approaches are ineffective without an additional knowledge source, such as a domain-specific ontology [15], [37], [6], [13] that is effort-intensive and costly to assemble. Our work addresses this issue without requiring such effort to structure the domain knowledge. To that end, we exploit the knowledge captured in Large Language Models (LLMs). LLMs are pre-trained with a vast amount of data covering a wide range of topics. They have been shown effective in answering questions regarding common sense [44], [7], [12] and domain-specific knowledge such as programming [31], [39], healthcare [10], [26], and law [38], [46]. LLMs are also capable of obtaining new knowledge from analogies and examples [24], [27]. Therefore, we propose to use LLMs for suggesting potential semantic relations among system capabilities subject to SLEEC requirements. For the example above, GPT-4.0 successfully suggested that `MuteNotifications` is contradictory with `SendUserWarning`.

On the other hand, using LLMs comes with its own risks, as they are notorious for producing significantly incorrect results [40], [47]. To prevent any falsely identified semantic relations from confounding WFI analysis, we also propose a

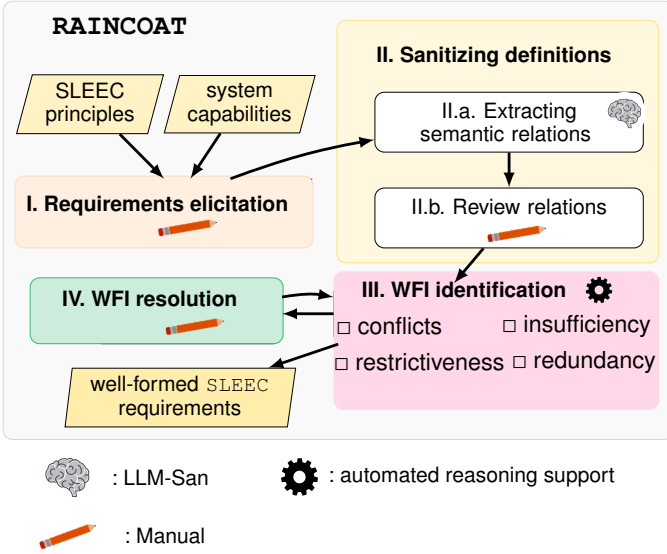


Fig. 1: Overview of RAINCOAT: the noRmAtive requirEmnts eliCitation and vAlidaTion approach.

lightweight logic-based filtering algorithm that incrementally extends relations via a set of inference rules and heuristically fixes logical inconsistencies on-the-fly. The filtered relations do not induce any logical fallacies and can be used directly for WFI analysis. We still, however, ask stakeholders to further filter the relations based on their domain-specific knowledge.

To summarise, the main contributions of our paper are: (1) An LLM-based technique that extracts semantic relations between the capabilities of a system under development, where these capabilities appear as non-terminal symbols in the SLEEC rules that encode the system’s N-NFRs; (2) The integration of our new semantic-relation extraction technique with existing automated reasoning methods, to develop a systematic end-to-end approach, RAINCOAT, for eliciting and analyzing normative requirements; (3) An extensive evaluation (with five non-technical stakeholders, using a range of real-world case studies) that shows the effectiveness and usability of our LLM-based technique and the RAINCOAT approach.

The rest of the paper is organized as follows: Sec. II gives the background material. Sec. III presents LLM-San, the new technique for determining semantic relations. Sec. IV gives an overview of our approach, RAINCOAT, for eliciting and analyzing normative requirements. Sec. V presents the evaluation of the effectiveness of LLM-San and RAINCOAT. Sec. VI overviews related work. Sec. VII concludes the paper and discusses future research directions.

II. PRELIMINARIES

In this section, we present background material. In Sec. II-A, we review the details of SLEEC, and in Sec. II-B, we present well-formedness issues of SLEEC requirements.

A. Normalized Representation of SLEEC

SLEEC is an event-based language for specifying normative requirements through the pattern “when trigger then response” [20]. It builds on propositional logic enriched with temporal

TABLE I: Syntax of normalized SLEEC with signature (E, M) .

Name	Definition
Term	$t := c : \mathbb{N} \mid m \in M \mid -t \mid t + t \mid c \times t$
Proposition	$p := \top \mid \perp \mid t = t \mid t \geq t \mid \text{not } p \mid p \text{ and } p \mid p \text{ or } p$
Obligation	$ob^+ := e \text{ within } t \mid ob^- := \text{not } e \text{ within } t$
Cond Obligation	$cob^+ := p \Rightarrow ob^+ \mid cob^- := p \Rightarrow ob^-$
Obligation Chain	$V_{cob} := cob \mid cob^+ \text{ otherwise } V_{cob}$
Rule	$r := \text{when } e \text{ and } p \text{ then } V_{cob}$
Fact	$f := \text{exists } e \text{ and } p \text{ while } (V_{cob} \mid \text{not } V_{cob})$

TABLE II: Capabilities, definitions and requirements of DAISY [9].

(a) DAISY capabilities.

- Meet a patient	- Issue commands to the patient	- Explain the protocols to the patient
- Examine the patient	- Collect patient data	- Perform triage assessment
- Provide instructions	- Repeat the instructions	- Request clinician intervention

(b) Sample DAISY SLEEC definitions.

event	ExamineState	measure	patientStressed
event	IdentifyDAISYTrust	measure	patientAgeConsidered
event	MeetingUser	measure	patientXReligion
event	MeetingPatient	measure	userDirectsOtherwise

(c) Sample DAISY SLEEC requirements.

Rule10	when MeetingUser then ExamineState
Rule13	when MeetingUser then IdentifyDAISYTrust
Rule16	when MeetingUser then ExaminingPatient within 30 minutes unless patientStressed
Rule19	when MeetingUser and patientXReligion then not ExaminingPatient unless userDirectsOtherwise unless medicalEmergency

constraints (e.g., **within x minutes**) and the constructs **unless**, specifying *defeaters*, and **otherwise**, specifying *fallbacks*. SLEEC is an expressive language, allowing specification of complex nested defeaters and responses. A SLEEC rule may be represented by multiple logically equivalent formulations, complicating analysis and reasoning. This issue can be addressed with the *normalized* SLEEC here, which simplifies nested response frameworks and removes defeaters, encoding their semantics using propositional logic. A translation from SLEEC to the normalized form is provided in [18].

Normalized SLEEC syntax. A normalised SLEEC *signature* is a tuple $S = (E, M)$, where $E = \{e_1, \dots, e_n\}$ and $M = \{m_1 \dots m_n\}$ are finite sets of symbols for *events* and *measures*, respectively. In the rest of the paper, we assume that all event symbols (e.g., OpenCurtain) are capitalized while measure symbols (e.g., underDressed) are not. Without loss of generality, we assume that every measure is numerical. The syntax of the normalized SLEEC can be found in Tbl. I. For example, an *obligation* $ob = h \text{ within } t$ specifies that the event e must (if $h = e$) or must not (if $h = \text{not } e$) occur within the time limit t . We omit the time limit t if $t = 0$.

Example 1. Consider DAISY [9], a sociotechnical AI-supported system to direct patients through an A&E triage pathway. Tbl. IIa describes DAISY capabilities. One normative rule (Rule16) is “**when** MeetingUser **then** ExaminingPatient **within 30 minutes unless** patientStressed”, where MeetingUser and ExaminingPatient are events and patientStressed a measure. Rule16 is normalized as “**when** MeetingUser **and not** patientStressed **then** ExaminingPatient **within 30 minutes**”.

Semantics. The semantics of SLEEC is described over *traces*, which are *finite* sequences of *states*

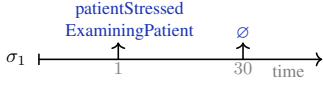


Fig. 2: Trace for the DAISY robot example.

$\sigma = (\mathcal{E}_1, \mathbb{M}_1, \delta_1), (\mathcal{E}_2, \mathbb{M}_2, \delta_2), \dots, (\mathcal{E}_n, \mathbb{M}_n, \delta_n)$. For every time point $i \in [1, n]$, (1) \mathcal{E}_i is a set of events that occur at i ; (2) $\mathbb{M}_i : M \rightarrow \mathbb{N}$ assigns every measure in M to a concrete value at time point i , and (3) $\delta_i : \mathbb{N}$ captures the value of time point i (e.g., the second time point can have the value 30, for 30 sec). We assume that the time values in the trace are strictly increasing (i.e., $\delta_i < \delta_{i+1}$ for every $i \in [1, n-1]$). Given a measure assignment \mathbb{M}_i and a term t , let $\mathbb{M}_i(t)$ denote the result of substituting every measure symbol m with $\mathbb{M}_i(m)$. Since $\mathbb{M}_i(t)$ does not contain free variables, the substitution results in a natural number. Similarly, given a proposition p , we say that $\mathbb{M}_i \models p$ if p is evaluated to \top after substituting every term t with $\mathbb{M}_i(t)$. Let a trace $\sigma = (\mathcal{E}_1, \mathbb{M}_1, \delta_1) \dots (\mathcal{E}_n, \mathbb{M}_n, \delta_n)$ be given.

(Positive obligation). A positive obligation e **within** t is fulfilled subject to time point i , denoted $\sigma \models_i e$ **within** t , if there is a time point $j \geq i$ such that $e \in \mathcal{E}_j$ and $\delta_j \in [\delta_i, \delta_i + \mathbb{M}_i(t)]$. That obligation is violated at time point j , denoted as $\sigma \not\models_i^j e$ **within** t , if $\delta_j = \delta_i + \mathbb{M}_i(t)$, and for every j' such that $j \geq j' \geq i$, e does not occur ($e \notin \mathcal{E}_{j'}$).

(Negative obligation). A negative obligation **not** e **within** t is fulfilled subject to time point i , denoted as $\sigma \models_i$ **not** e **within** t , if for every time point j such that $\delta_j \in [\delta_i, \delta_i + \mathbb{M}_i(t)]$, $e \notin \mathcal{E}_j$. The negative obligation is violated at time point j , denoted as $\sigma \not\models_i^j$ **not** e **within** t , if (1) $\delta_j \in [\delta_i, \delta_i + \mathbb{M}_i(t)]$, (2) e occurs ($e \in \mathcal{E}_j$), and (3) for every $j \geq j' \geq i$, e does not occur ($e \notin \mathcal{E}_{j'}$).

(Conditional obligation). A conditional obligation $p \Rightarrow ob$ is fulfilled subject to time point i , denoted as $\sigma \models_i (p \Rightarrow ob)$, if p does not hold at time point i (that is, $\mathbb{M}_i(p) = \perp$) or the obligation is fulfilled ($\sigma \models_i ob$). Moreover, $p \Rightarrow ob$ is violated at time point j , denoted as $\sigma \not\models_i^j (p \Rightarrow ob)$, if p holds at i ($\mathbb{M}_i(p) = \top$) and ob is violated at j ($\sigma \not\models_i^j ob$).

(Obligation chain). A chain $cob_1^+ \text{ otherwise } cob_2^+ \dots cob_m$ is fulfilled subject to time point i , denoted as $\sigma \models_i cob_1^+ \text{ otherwise } cob_2^+ \dots cob_m$, if (1) the first obligation is fulfilled ($\sigma \models_i cob_1^+$) or (2) there exists a time point $j \geq i$ such that cob_1^+ is violated (i.e., $\sigma \not\models_i^j cob_1$) and the rest of the obligation chain (if not empty) is fulfilled at time point j ($\sigma \models_j cob_2^+ \text{ otherwise } \dots cob_m$).

(Rule). A rule **when** e **and** p **then** \bigvee_{cob} is fulfilled in σ , denoted as $\sigma \models$ **when** e **and** p **then** \bigvee_{cob} , if for every time point i , where event e occurs ($e \in \mathcal{E}_i$) and p holds ($\mathbb{M}_i(p) = \top$), the obligation chain is fulfilled subject to time point i ($\sigma \models_i \bigvee_{cob}$). A trace fulfills a rule set $Rules$, denoted as $\sigma \models Rules$, if it fulfills every rule in the set (i.e., for every $r \in Rules$, $\sigma \models r$).

Example 2. Consider the normalized SLEEC rule `Rule16` in Example 1 and the trace σ_1 shown in Fig. 2, corresponding to $(\mathcal{E}_1, \mathbb{M}_1, \delta_1)$, where $\mathcal{E}_1 = \text{MeetingUser}$, $\mathbb{M}_1(\text{patientStressed})$

$= \top$, and $\delta_1 = 1$. The trace σ_1 fulfills the rule `Rule16` (i.e., $\sigma_1 \models \text{patientStressed}$), because the rule's triggering condition **not** `patientStressed` is not satisfied, i.e., the patient is `patientStressed` at the time of the meeting).

Definition 1 (Behaviour defined by Rules). Let a rule set $Rules$ be given. The accepted *behaviour* defined by $Rules$, denoted as $\mathcal{L}(Rules)$, is the largest set of traces such that every trace σ in $\mathcal{L}(Rules)$ respects $Rules$, i.e., $\sigma \in \mathcal{L}(Rules)$, $\sigma \models Rules$.

On top of the usual rules, SLEEC also uses *Facts* as an utility construct describing the test scenarios for rules. A *fact* (see Tab. I) **exists** $e \wedge p$ **while** $(\bigvee_{cob} \mid \text{not } \bigvee_{cob})$ asserts the existence of an event e under condition p while an obligation chain \bigvee_{cob} is either satisfied or violated. Facts can be used to describe undesirable behaviors, denoted as *concerns*, which should be blocked by SLEEC rules. Alternatively, Facts can also be used to describe functional goals, denoted as *purposes*, which should be allowed by SLEEC rules.

B. Well-Formedness Issues (WFI)

SLEEC Rules might be subject to well-formedness issues (WFIs): *vacuous conflict*, *redundancy*, *insufficiency*, *over-restrictiveness* and *conditional conflict*. We explain each WFI briefly and refer to [17] for a formal definition.

Vacuous conflict. A rule is *vacuously conflicting* in a rule set $Rules$ if triggering this rule always leads to a conflict with some other rules in $Rules$.

Situational conflict. A rule is *situational conflicting* in a rule set $Rules$ if triggering this rule under a certain situation which defines a history of events and measures, always leads to a conflict with some other rules in the future.

Redundancy. A rule is *redundant* in a rule set $Rules$ if this rule is logically implied by other rules in $Rules$.

Insufficiency. A rule set $Rules$ is *insufficient* subject to a given concern if some behavior represented by the concern is realizable while respecting $Rules$.

Over-restrictiveness. A rule set $Rules$ is *overly restrictive* subject to a given purpose if none of the functional goals represented by the purpose is realizable while respecting $Rules$.

WFIs impose threats to the validity of SLEEC rules. Conflicting rules must be resolved to avoid inevitable SLEEC harms. Insufficient and overly-restrictive rule sets also need to be fixed, to avoid executions of undesirable behaviors and violations of functional goals, respectively. Situational conflicts and redundancies also require attention. The former capture dead-end situations where conflicts become inevitable and the latter complicates management of evolving requirements.

Order of Resolving WFIs. Different types of WFIs have different levels of severity and might depend on each other.

Vacuous conflicts should be resolved prior to situational conflicts because any situation that triggers a vacuously conflicting rule will also create a conflicting situation. For example, consider $R_1 = \text{when } A \text{ then } B$. This rule is vacuously conflicting due to another rule: $R_2 = \text{when } A \text{ then not } B \text{ within}$

Redundant SLEEC rule: Rule10 when MeetingUser then ExamineState
Because of the following SLEEC rule:
Rule10_01 when MeetingUser then IntentExaminePatient
Rule10_02 when IntentExaminePatient then ExamineState

Fig. 3: Redundancy diagnosis example.

10 minutes. This implies that $R1$ is also situationally conflicting since any situation that triggers $R1$ leads to a conflict. Insufficiency and over-restrictiveness are checked and resolved after both types of conflict because the behaviors defined by conflicting rules are inconsistent, and thus might affect the judgment of insufficiency or over-restrictiveness analysis. For example, without first resolving the vacuous or situational conflict between $R1$ and $R2$, the conflicting situations are excluded from the insufficiency and over-restrictiveness analysis. The concern **exists** A **while** X can never be identified in any currently conflicting situation (and cannot be identified at all if a vacuous conflict exists), thus it is blocked by the conflict.

Redundancies are checked last because they are the least severe WFIs, and because they are affected by conflicts. For example, without resolving the vacuous conflict between $R1$ and $R2$, the rule $R3 = \text{when } A \text{ then } C$ is redundant because triggering $R3$ would also trigger $R1$ and subsequently cause a conflict. Moreover, the conflict subsumes the response of $R3$ (e.g., c), thus making the rule redundant.

LEGOS-SLEEC. [17] was proposed to identify and resolve WFIs. It encodes the semantics of rules into FOL* and turns WFI conditions into FOL* satisfiability queries. The satisfiability results, either satisfying solutions or proofs of unsatisfiability, are then used to diagnose the causes of WFIs.

Example 3. Consider a rule set $\text{Rules} = \{\text{Rule10}, \text{Rule10_01}, \text{Rule10_02}\}$, where $\text{Rule10}, \text{Rule10_01}, \text{Rule10_02}$ are rules in Tbl. IIc. To check whether Rule10 is redundant in Rules , we can use LEGOS-SLEEC which checks the satisfiability of the query $\{\neg \text{Rule10}\} \cup \{r \mid r \in \text{Rule10} \setminus \{\text{Rule10}\}\}$. The query is unsatisfiable (UNSAT), which means that Rule10 is a logical consequence of the $\{\neg \text{Rule10}\} \cup \{r \mid r \in \text{Rule10}\}$; hence, Rule10 is redundant. LEGOS-SLEEC provides a diagnosis shown in Fig. 3, with the reason of redundancy highlighted.

III. SANITIZING DEFINITIONS

The semantics of SLEEC (see Sec.II-A) assumes that the capabilities of the systems (represented by events and measures) are independent. SLEEC does not support the description of the relations between these capabilities. In this section, we first propose an extension to SLEEC for describing the relations between capabilities, denoted as CR (Sec.III-A). Then, we present an LLM-aided approach to automatically extract CRs from the textual semantics of the capabilities (Sec. III-B).

A. Capability Relations (CR)

We propose extending SLEEC to capture three types of binary relations: (1) between two events, (2) between two measures, and (3) between an event and a measure.

Relations between events. We capture the following list of binary relations between events e_a and e_b :

- e_a **hypernym** e_b iff (if and only if) for every state, the occurrence of e_a implies the occurrence of e_b .
- e_a **isContradictoryWith** e_b iff for every state, e_a and e_b never occur simultaneously.
- e_a **happensBefore** e_b iff for every state s_1 where e_a occurs, there exists some state s_2 prior to s_1 in which e_b occurred.
- e_a **equal** e_b iff e_a **hypernym** e_b and e_b **hypernym** e_a .

The formal definitions of these relations are presented in the supplementary material [2]. Table III illustrates each relations.

Relations between measures. We capture the following list of binary relations between propositions p_a and p_b over measures:

- p_a **imply** p_b iff for every state, p_a holding implies that p_b also holds.
- p_a **mutuallyExclusive** p_b iff for every state, p_a and p_b never hold simultaneously.
- p_a **opposite** p_b iff for every state, p_a holds if and only if p_b does not hold.
- p_a **equal** p_b iff p_a **imply** p_b and p_b **imply** p_a .

Formal definitions of these relations are in the supplementary material [2]. Tbl. IV illustrates each relation type.

Relations between events and measures. We capture the following list of relations between events and measures:

- p **forbids** e iff for every state, if p holds, it implies that the occurrence of e is not possible at those times.
- e **induces** p iff for every state, the occurrence of e implies that p holds at the same time.
- **when** e_a **then** p **until** e_b iff p holds for all states starting from a state where e_a occurs until a state where e_b occurs.
- **when** e **then** p **for** t iff p holds for all states starting from a state s_1 where e occurs until a state s_2 such that s_2 is t units of time after s_1 .

The formal definitions of the above relations are presented in the supplementary material [2]. Table V gives examples of each type of relation. Note that **forbids** and **induces** describe relations that are concerned with the same state (i.e., with the same time point), while **when ... then ... until ...** and **when ... then ... for ...** are temporal relations between events and measures across multiple states.

Rationale behind the selection of semantic relation terms.

The terms used to describe the semantic relations between events and measures were determined through collaboration with a non-technical stakeholder, a philosopher with expertise in common-sense knowledge and intuition-based reasoning. The relations were chosen to be easy to understand and intuitive where possible. The majority of the terms have everyday usage across multiple contexts and domains, thereby

TABLE III: Example relations between events.

Relation	SLEEC DSL example
hypernym	<i>water is an instance of liquid</i> DrinkWater hypernym DrinkLiquid
equal	patient and client are interchangeable, when patients are the only clients CallPatient equal CallClients
contradictory	<i>impossible to open and close the door simultaneously</i> OpeningDoor isContradictoryWith ClosingDoor
happen before	<i>locking the door occurs after closing it</i> ClosingDoor happensBefore LockingDoor

TABLE IV: Example semantic relations between measures.

Notation	SLEEC DSL example
imply	<i>door opened can not be locked</i> doorOpened imply not doorLocked
mutual exclusive	<i>door opened can not be locked</i> doorOpened mutuallyExclusive doorLocked
opposite	<i>the door can be either opened or closed</i> doorOpened oppositeTo doorClosed
equal	<i>patient and client are interchangeable, when patients are the only clients</i> patientConsented equal userConsented

making them easier for stakeholders to adopt. These ‘common sense’ relations include ‘happens before’, ‘equal’, ‘imply’, ‘opposite’, ‘forbids’, ‘contradictory’ and ‘induces’. Stakeholders will likely already possess an understanding of these relations prior to being introduced to them in the context of events and measures (although each relation is still thoroughly explained to stakeholders and illustrated with examples to aid comprehension). In the absence of readily apparent intuitive or common-sense equivalents and to avoid verbose descriptions, terms were selected that instead capture the intended semantic relation. ‘Hypernym’ was selected based on its frequent use in semantic relation databases such as WordNet [35], [34], [42]. The term ‘mutually exclusive’ was chosen due to its frequent use in philosophy, law, computer science and other relevant disciplines. To ensure concurrence across disciplines, the precise meaning intended for this work was specified and standardised (as was the case for each relation identified).

B. LLM-aided CR Extractions

Given sets of symbols E and M for the events and measures of a SLEEC document, the algorithm LLM-San (see Alg. 1) computes potential binary relations between the event and measure symbols in E and M according to their textual semantics. LLM-San first prepares and executes queries (Alg. 1 LL:1-2) to an LLM to extract a set of candidate relations (as presented in Sec.III-B1). Then, it filters (L:3) the candidate set to ensure logical consistency (as presented in Sec. III-B2).

Remark 1. LLM-San automatically extracts relations only between a pair of events or measures. Relations between an event (and multiple events, in the case of **when until**) and a measure should be more carefully defined by the users.

1) *Relation discovery:* LLM-San initially queries an LLM to discover a set of binary relations between two events or two measures. The query to the LLM includes the following information: (1) the grammar of SLEEC (shown in Table I); (2) the set of event and measure symbols in E and M , respectively; (3) the textual definition and an illustration for each type of relation. For example, the query for the relation

TABLE V: Example dependencies between events and measures.

ID	SLEEC DSL example
induce	CollectConsent induces consentObtained
forbidden	inWater forbids CarStartSpeeding
until	when CollectConsent then consentObtained until ConsentWithdraw
for	when LoginConfirmed then loggedIn for 10 minutes

Algorithm 1 LLM-San (E, M)

Require: E and M are the complete sets of event and measure symbols, respectively
Ensure: Rel_f is a consistent set of relations between events and measures in E and M , respectively.
1: $query \leftarrow \text{PREPAREQUERY}(E, M)$
2: $Rel \leftarrow \text{LLM}(query)$
3: $Rel_f \leftarrow \text{Filter}(Rel)$
4: **return** Rel_f

happensBefore is shown in Fig.5; (4) the JSON format for the expected output. For example, the output format for a discovered **happensBefore** relation is shown in Fig. 6.

LLM-San sends a query to the LLM and then waits for it to populate a JSON file containing a set of discovered relations and their verdicts $Rel_s = r_1, \dots, r_n$. A relation is considered positive if it is confirmed by the LLM (e.g., e_a **happensBefore** e_b) and negative if it is rejected (e.g., **not** (e_a **happensBefore** e_b)). Fig. 7 demonstrates the LLM’s output representing a positive relation of **happensBefore**. For all relations, LLM-San asks the LLM to provide a justification before producing the final verdict. This simulates the forward-thinking process and prevents the LLM from retrofitting a justification for a random verdict.

2) *Filtering relations via logic inference:* The initial set of candidate relations Rel produced by the LLM might contain semantic errors that could lead to logical inconsistencies and subsequently confound the WFIs analysis. While LLM-San, as an automated procedure, cannot find all semantic errors in Rel , it addresses any potential logical inconsistencies in it.

LLM-San calls Alg. 2 on Rel to remove candidate relations that could lead to logical inconsistencies, based on the inference rules shown in Fig. 4. These inference rules derive new relations from existing ones in Rel . Let Rel^* be the fixed-point set of all relations that are either in Rel or derivable from Rel^* . A relation r witnesses an inconsistency if Rel^* contains both the positive relation r and the negative relation **not** (r). A relation set Rel is *consistent* if Rel^* has no inconsistency witnesses. Alg. 2 takes as argument a set Rel of relations and returns a candidate set $Rel_f \subseteq Rel$ that is consistent.

All inference rules in Fig. 4 are Horn clauses in the form $r_1, r_2, \dots, r_n \Rightarrow r_{cons} \mid \text{not}(r_{cons})$. Only positive relations in Rel can lead to the deduction of new relations using the rules. Negative relations are used for identifying inconsistency witnesses. Therefore, an empty set of relations is *vacuously consistent*. It is, however, not very useful. LLM-San aims to greedily find a reasonably large subset of Rel that is consistent. It is *not guaranteed*, however, that LLM-San returns the largest consistent subset, as the problem of finding this subset is generally NP-hard, even for Horn logic [29].

Given an input set Rel , Alg. 2 iteratively computes Rel^*

Inference rules for relations between two events.			
$\frac{e_a \text{ isContradictoryWith } e_b}{\text{not } (e_a \text{ hypernym } e_b)} (IP1^-)$	$\frac{e_a \text{ happensBefore } e_b}{\text{not } (e_a \text{ hypernym } e_b)} (IP2^-)$	$\frac{e_a \text{ hypernym } e_b \wedge e_b \text{ hypernym } e_c}{e_a \text{ hypernym } e_c} (IPtrans^+)$	$\frac{e_a \text{ equal } e_b}{e_a \text{ hypernym } e_b} (IPEQ^+)$
$\frac{e_a \text{ hypernym } e_b}{\text{not } (e_a \text{ isContradictoryWith } e_b)} (ME1^-)$	$\frac{e_b \text{ isContradictoryWith } e_a}{e_a \text{ isContradictoryWith } e_b} (MEcomm^+)$	$\frac{e_a \text{ hypernym } e_b \wedge e_b \text{ isContradictoryWith } e_c}{e_a \text{ isContradictoryWith } e_c} (MEtrans^+)$	
	$\frac{e_a \text{ hypernym } e_b \wedge e_b \text{ hypernym } e_a}{e_a \text{ equal } e_b} (EQIP^+)$	$\frac{e_a \text{ equal } e_b}{e_b \text{ equal } e_a} (EQcom^+)$	
	$\frac{e_a \text{ happensBefore } e_b \wedge e_b \text{ happensBefore } e_c}{e_a \text{ happensBefore } e_c} (HBtrans1^+)$	$\frac{e_a \text{ happensBefore } e_b \wedge e_c \text{ hypernym } e_b}{e_a \text{ happensBefore } e_c} (HBtrans2^+)$	
	$\frac{e_b \text{ hypernym } e_a \wedge e_b \text{ happensBefore } e_c}{e_a \text{ happensBefore } e_c} (HBtrans3^+)$		
Inference rules for relations between two propositions over measures.			
$\frac{p_a \text{ mutualExc } p_b}{\text{not } (p_a \text{ imply } p_b)} (MIP1^-)$	$\frac{p_a \text{ imply } p_b \wedge p_b \text{ imply } e_c}{p_a \text{ imply } p_c} (MIPtrans^+)$	$\frac{p_a \text{ equal } p_b}{p_a \text{ imply } p_b} (IPEQ1^+)$	$\frac{p_a \text{ equal } p_b}{\neg p_a \text{ imply } \neg p_b} (IPEQ2^+)$
$\frac{p_a \text{ imply } p_b}{\text{not } (p_a \text{ mutualExc } p_b)} (MME1^+)$	$\frac{p_b \text{ mutualExc } p_a}{p_a \text{ mutualExc } p_b} (MMEcomm^+)$	$\frac{p_a \text{ imply } p_b \wedge p_b \text{ mutualExc } p_c}{p_a \text{ mutualExc } p_c} (MMEtrans^+)$	
$\frac{p_a \text{ imply } p_b \wedge p_b \text{ imply } p_a}{p_a \text{ equal } p_b} (MEQIP^+)$	$\frac{p_a \text{ opposite } \neg p_b \wedge p_b \text{ imply } p_a}{p_a \text{ equal } p_b} (MEQOP^+)$	$\frac{p_a \text{ equal } p_b}{p_b \text{ equal } p_a} (MEQcom^+)$	
$\frac{p_a \text{ equal } \neg p_b}{p_a \text{ opposite } p_c} (MOPEQ^+)$	$\frac{p_b \text{ opposite } p_s}{p_a \text{ opposite } p_b} (MOPcoms^+)$	$\frac{p_a \text{ mutualExc } p_b \wedge \neg p_a \text{ mutualExc } \neg p_b}{p_a \text{ opposite } p_b} (MOPME^+)$	

Fig. 4: Horn inference rules for binary relations between two events or two measures. Rules with trailing “+”s and “-”s derive positive and negative relations, respectively. Alg. 2 always propagates on negative (“-”) before propagating on positive (“+”) rules.

Algorithm 2 CheckConsistency(*Rel*)

Require: *Rel* is a set of binary relations between two events or two measures.
Ensure: *Rel_f* is a subset of *Rel*, and is logically consistent w.r.t. the inference rules in Fig. 4.

- 1: *Rel** ← *Rel*, *Rel'* ← ∅
- 2: **while** ⊤ **do**
- 3: *Rel** ← *Rel* * ∪ *Rel'*
- 4: *Rel** ← {*r* | *r* is negative ∨ **not** (*r*) ∉ *Rel**}
- 5: *old_Rel** ← *Rel**
- 6: *Rel** ← APPLYRULES⁻(*Rel**)
- 7: *Rel**, *Rel'*? ← APPLYRULES⁺(*Rel**)
- 8: **if** *old_Rel** = *Rel** ∧ *Rel'*? = ∅ **then**
- 9: **return** *Rel* ∩ *Rel**
- 10: **end if**
- 11: *Rel'* ← LLM(*Rel'*?)
- 12: **end while**

For every pair of events e_a and e_b , can you please answer the following question: Is it the case that there is always an occurrence of e_a that happens before every occurrence of e_b ? If yes, please say e_a **happensBefore** e_b . If no, please say **not** (e_a **happensBefore** e_b). For example, `CreateForm` **happensBefore** `ShowForm` is correct because `CreateForm` is the prerequisite of `ShowForm`.

Fig. 5: LLM query for finding binary relations of “**happensBefore**”.

from *Rel* (L:1) while fixing any identified inconsistency on-the-fly. For each iteration, before applying inference rules, LLM-San first checks *Rel** for local consistency (L: 4) and overrides all inconsistent positive relations $r \in Rel^*$ (i.e., if **not** (r) ∈ *Rel**). Negative relations are favored over positive ones to speed-up convergence since all inference rules are Horn. After fixing local consistencies, all rules for deriving negative relations (i.e., inference rules whose name ends with “-”) are exhaustively applied (L:6). When an inference rule derives a negative relation **not** (r_{cons}), there are three cases:

```
"Relationship": {
  "event1": {name of EVENT A},
  "event2": {name of EVENT B},
  "justification": {your justification},
  "HappensBefore": {true or false}
}
```

Fig. 6: Output format for binary relations of “**happensBefore**”.

```
"Relationship": {
  "event1": "PreparingDeployment",
  "event2": "AgentDeployed",
  "justification": "An agent must be prepared for
                  deployment before it can be deployed.",
  "HappensBefore": true
}
```

Fig. 7: LLM output for a positive relation of “**happensBefore**”.

(1) **not** (r_{cons}) ∈ *Rel**, (2) r_{cons} ∈ *Rel**, and (3) otherwise. In the first case, the consequence is consistent with *Rel** and no action is required. In the second case, there is an inconsistency in *Rel** witnessed by r_{cons} . Alg. 2 then resolves the inconsistency by updating r_{cons} with **not** (r_{cons}) in *Rel**. In the third case, **not** (r_{cons}) is added to *Rel**.

Example 4. Let $Rel = \{r1 = e_a \text{ hypernym } e_b, r2 = e_a \text{ happensBefore } e_b\}$ be given. Alg. 2 first derives a relation $r3 = \text{not } (e_a \text{ mutualExc } e_b)$ from $r1$ and $r2$ by rule “*IP2*-”. The derived relation $r3$ conflicts with $r1$ in *Rel* (and *Rel**). The conflict is resolved by replacing $r1$ with $r3$ in *Rel**.

After exhausting all negative inference rules and obtaining a new *Rel** (L:6), Alg. 2 then proceeds to apply the positive rules (L:7). If an inference rule derives a positive relation r_{cons} , there are three cases: (1) r_{cons} ∈ *Rel**, (2) **not** (r_{cons}) ∈ *Rel**, and (3) otherwise. In the first case, the consequence is

```
Please trying to fill in the content consider the context given above:
{'Relationship': {'event1': 'UserWantsToCook',
                 'event2': 'BeforeCookingBegins',
                 'justification': '{fill in your justification}',
                 'HappensBefore': '{fill in here, true of false}'}}
```

Fig. 8: Follow-up query for confirming the relation `UserWantsToCook happensBefore BeforeCookingBegins`.

consistent. In the second case, there is an inconsistency in Rel^* witnessed by r_{cons} . Alg.2 resolves this inconsistency by updating one of the premises $r_i \in r_1, \dots, r_n$ with $\neg r_i$ in Rel^* . In the third case, instead of directly adding r_{cons} to Rel^* , Alg. 2 prepares a follow-up query to the LLM about r_{cons} for confirmation. Note that this follow-up query is necessary because the LLM did not initially discover r_{cons} as a positive relation in Rel . This might imply that **not** (r_{cons}) $\in Rel$ if the LLM follows the closed-world assumption, or ‘unknown’ otherwise. We store all follow-up queries to the LLM in a set $Rel?$ when applying the positive inference rules. An example of the follow-up query is shown in Fig. 8.

After exhausting the positive inference rules, the follow-up queries $Rel?$ are sent to the LLM (L:11). The results, denoted as Rel' , comprise a set of positive and negative relations, representing confirmations and rejections of the queries in $Rel?$, respectively. Alg. 2 then merges Rel^* with Rel' (L:3), and starts a new iteration of inferences (LL:6-11). Alg. 2 terminates when Rel^* reaches a fixed point and no follow-up queries are produced (L:8). Finally, Alg. 2 returns the intersection of Rel^* and the original relations Rel (L:9) as the consistent set of relations. Note that not every relation in Rel^* can be derived from the final return set because some relations might be updated after being used to derive other relations in Rel^* . We do not include these non-derivable relations in the return even though including them does not affect consistency. On the other hand, the derivable relations in Rel^* do not need to be in the final return set since they can be logically inferred, and thus are redundant.

Example 5. Let $Rel = \{r1 = e_a \text{ **hypernym** } e_b, r2 = e_b \text{ **isContradictoryWith** } e_c, r3 = e_a \text{ **hypernym** } e_c\}$ be given. Running Alg. 2 first derives a positive relation $r4 = e_a \text{ **mutualExc** } e_c$ from $r1$ and $r2$ by rule “*METrans+*”. $r4$ and **not** ($r4$) are not in Rel^* , and $r4$ is added to $Rel?$ to be queried by the LLM. Suppose the LLM confirms $r4$ and adds it to Rel^* . Then in the next iteration, Alg. 2 derives $r5 = \text{not } (e_a \text{ **hypernym** } e_c)$ from $r4$ by rule “*IP1-*”, which conflicts with $r1$ in Rel . The conflict is resolved by replacing $r1$ with $r5$ in Rel^* . At this point, $Rel^* = \{r2, r3, r4, r5\}$ is a consistent set of relations. Alg. 2 returns the intersection of Rel and Rel^* , which is $\{r2, r3\}$.

Remark 2. Alg. 2 terminates. This is because (1) there is a finite number of possible binary relations, and (2) every relation r in Rel^* can change at most once: a positive relation r can be updated to $\neg r$, but $\neg r$ can never be updated to r .

LLM-San returns the consistent relation set Rel_f as the candidate relations. Stakeholders are expected to review these

candidates and validate the correct ones, which are then automatically included in the SLEEC rules for WFIs analysis.

IV. THE RAINCOAT APPROACH

In this section, we introduce our tool-supported approach, RAINCOAT, for normative requirements elicitation and validation, depicted in Fig. 1. The approach consists of four distinct stages which we discuss below.

I. Initial requirements elicitation. The goal of this stage is to systematically identify *preliminary normative requirements* and obtain their *formal* and *machine-readable* representations.

RAINCOAT follows the approach presented in [45] to first contextualize the high-level SLEEC principles by mapping them onto the agent capabilities. Consider DAISY, the robotics system described in Sec. II. Its capabilities to “meet a patient” and “explain the interaction protocols to the patient” are mapped to the SLEEC principles “autonomy” and “self-determination”, and the mapping could result in the following preliminary requirement: “when DAISY meets a patient, it shall explain the interaction protocol to her”. Note that during the elicitation process, new capabilities might be recommended by non-technical stakeholders if they believe that a SLEEC harm could occur otherwise. For instance, to comply with privacy legislation, stakeholders could recommend adding the “RemoveData” capability to DAISY, which grants users the right to be forgotten.

For formalizing preliminary requirements, RAINCOAT follows the approach presented in [17], which uses the domain-specific language SLEEC. This DSL has been shown to be accessible to stakeholders from different fields, including lawyers, philosophers, ethicists, roboticists, and software engineers [20], [21], [16], [17]. For example, the preliminary autonomy requirement is formalized as a SLEEC rule `when MeetingPatient then ExplainProtocols`. More examples of DAISY’s SLEEC requirements and their definitions are provided in Tbl. IIc and IIb, respectively.

II. Sanitizing definitions. This stage aims to enrich the preliminary set of normative requirements by capturing and integrating the semantic relations (see Sec. III-A) between system capabilities (i.e., events and measures). This stage uses our LLM-aided CR extraction algorithm LLM-San (see Sec. III-B) to compute a set of candidate relations which are then (manually) validated by the stakeholders. The validated relations are integrated into the preliminary normative requirements. For example, the relation `MeetingUser equal MeetingPatient` has been suggested, validated and integrated for DAISY because patients are the only DAISY users.

III. Identification of well-formedness issues. In this stage, we detect WFIs (see Sec.II-B) in the preliminary normative requirements using the existing automated reasoning technique, LEGOS-SLEEC, in the order (1) *vacuous conflicts*, (2) *situational conflicts*, (3) *insufficiencies & over-restrictiveness*, and (4) *redundancies*. For each issue, we provide a *diagnosis* to help stakeholders understand exactly which rules and clauses caused the WFI. Consider a rule set containing

Conflicting SLEEC rule: Rule10 when MeetingUser then ExamineState

Because of the following SLEEC rule:
isContradictoryWith MeetingUser ExamineState

Fig. 9: A diagnosis showing that Rule10 and the relation MeetingUser **isContradictoryWith** with ExamineState cause a vacuous conflict.

Rule10 and a relation MeetingUser **isContradictoryWith** with ExamineState. A vacuous conflict is identified and a diagnosis (see Fig. 9) is produced to show that Rule10 and the relation are the causes of the conflict.

If an WFI is identified, RAINCOAT jumps to the (manual) ‘WFI Resolution’ stage (Stage IV). The process terminates when no further WFIs have been identified, resulting in well-formed SLEEC requirements.

IV. WFI Resolution. The objective in this stage is for the user to use the provided diagnosis to (manually) address the identified WFIs. For example, consider resolving the vacuous conflict given the diagnosis in Fig. 9. We can deduce that the rule should not require a simultaneous response. Instead, it could be rewritten to include a time delay, such as: **when MeetingUser then ExamineState within 10 minutes**. Stakeholders are expected to resolve each issue with the exception of redundancies, where they may choose to intentionally preserve some of them. After resolving the WFIs, the stakeholders return to Stage III to confirm that the refinement process did not introduce new WFIs. If the resolution process involved updating some of the requirements definitions, the stakeholders should return to Stage II to capture the potential dependencies in the updated capabilities.

RAINCOAT actively engages stakeholders in the elicitation and rule analysis process, and completes when the stakeholders have resolved all conflicts and concerns, and deem the elicited requirements to be comprehensive. Our experiments (see Sec. V) show that the process successfully terminates in 1-3 iterations.

V. EVALUATION

In this section, we evaluate the effectiveness of both the LLM-San technique, which infers the relation between the capabilities in the SLEEC-DSL rules, and the overall approach, RAINCOAT, for supporting non-technical stakeholders for eliciting a comprehensive and well-formed set of normative requirements. LLM-San implementation and the evaluation artifacts are available in [2].

A. Effectiveness of LLM-San

We aim to answer: **RQ1**: What is the effectiveness of using LLM-San w.r.t. the number of relevant and spurious WFIs identified? Through answering this question, we aim to determine whether the resulting set of requirements is more sound and comprehensive.

Models and Methods. Non-technical stakeholders (N-TSs) were asked to review the capabilities relation inferred by LLM-San on nine real-world case-studies. The N-TS group included an ethicist, a lawyer, a philosopher, and a psychologist. The case studies were taken from the repository of normative requirements [17]: (1) ALMI [25]: a system assisting elderly or disabled users in a monitoring/advisory role and with everyday tasks; (2) ASPEN [11]: an autonomous agent dedicated to forest protection, providing both diagnosis and treatment of various tree diseases; (3) AutoCar [4]: a system that implements emergency-vehicle priority awareness for autonomous vehicles; (4) BSN [23]: a healthcare system detecting emergencies by continuously monitoring the patient’s health status; (5) DressAssist [30], [45]: an assistive and supportive system used to dress and provide basic care for those in need; (6) CSI-Cobot [43]: a system ensuring the safe integration of industrial collaborative robot manipulators; (7) DAISY [9]: a sociotechnical AI-supported system that directs patients through an A&E triage pathway (our running example); (8) DPA [1]: a system to check compliance of data processing agreements against the General Data Protection Regulation; (9) SafeSCAD [8]: a driver attentiveness management system to support safe shared control of autonomous vehicles. We have selected these case studies because they have been used in existing work [17] which examined them for well-formedness issues. However, the authors of [17] assumed independence of all capabilities referenced in the rules. Our goal is to determine whether capturing the relations between capabilities is important in practice.

For each case study, two to four N-TSs individually labelled the semantic relations inferred by LLM-San as correct or incorrect, with a justification if needed, and timed this process. After capturing the relations classified as correct, we analyzed the case studies for any new well-formedness issues (WFI). Subsequently, we requested that stakeholders review the newly-identified WFIs to ensure that they are *relevant*, i.e., not spurious.

Results. We consider semantic relations to be *correct* (TP) if they were classified as correct by the majority of the N-TSs. The others are considered *spurious* (FP). For each case study, we report the type of semantic relation that has been captured as well as the new relevant well-formedness issues identified. The results are shown in Tbl. VI.

LLM-San inferred the total of 103 semantic relations for all but one (DressAssist) case studies, and the N-TSs classified 53 of them as correct. The classification took between 4 minutes (BSN) and 30 minutes (DPA). The number of correct relations added per case study ranged between 19 (ASPEN) and one (BSN and SafeSCAD). The majority of the added relations (27) involved contradicting events, that is, there were 27 pairs of events whose expected effects were mutually exclusive. The second most common relation type added (13) was the one involving mutual exclusiveness between two measures. We observed that a higher number of events and measures did not lead to a proportionally higher number of relations between

TABLE VI: LLM-San effectiveness. We record: true positive (TP) and false negative (FP) in the extracted capabilities relation (rel.); the number of hyponyms (hyp.), coincidences (coi.), contradicting (cont.), happening before (h.b.), implications (imp.), mutually exclusive (m.e.), equivalences (eq.), induce (induc), and forbid (forb.) relations added; the number of new issues identified (after capturing the semantic relations) for vacuous conflict (v-conf.), situational conflict (s-conf.), redundancy (redund.), restrictiveness (restrict.), and insufficiency (insuff.).

case studies	rules (evnt. – msr.)	relations (TP - FP)	number of relations by type								new WFI				
			hyp.	coi.	cont.	h.b.	imp.	m.e.	eq.	induc.	forb.	v-conf.	s-conf.	redund.	insuffi.
ALMI	39 (41 – 15)	5 - 2	0	0	0	0	2	2	1	0	0	0	0	0	0
ASPEN	15 (25 – 18)	19 - 8	0	0	7	1	2	7	1	0	1	1	0	1	0
AutoCar	19 (36-26)	6 - 19	0	0	5	0	0	0	0	0	1	0	0	0	0
BSN	29 (33-31)	1 - 3	0	0	1	0	0	0	0	0	0	0	0	2	0
CSI-Cobot	20 (23-11)	3 - 3	0	0	3	0	0	0	0	0	0	0	0	0	0
DAISY	26 (45-31)	15 - 5	0	0	9	0	0	3	0	0	3	5	0	4	0
DPA	26 (28-25)	3 - 2	0	0	2	0	0	1	0	0	0	0	0	0	0
DressAssist	31 (54-42)	0 - 0	0	0	0	0	0	0	0	0	0	0	0	0	0
SafeSCAD	28 (29-20)	1 - 8	0	0	0	0	1	0	0	0	0	0	0	0	0

them. We identified 13 new well-formedness issues, all of which have been classified as relevant by the N-TSs. Out of them, six were instances of vacuous conflicting rules, meaning that triggering them would inevitably lead to a conflict! The remaining seven instances represented redundant requirements. To summarize, capturing semantic relations using LLM-San enabled us to identify new relevant WFIs, answering **RQ1**.

B. Effectiveness of RAINCOAT

We now aim to evaluate the effectiveness of the overall approach, RAINCOAT, in eliciting a comprehensive and well-formed set of normative requirements. Specifically, We aim to assess whether the use of RAINCOAT facilitates faster convergence and yields higher quality requirements by answering **RQ2**: How does RAINCOAT impact the overall elicitation and validation process in terms of the number of iterations, convergence speed, and quality compared to common practices of non-technical stakeholders?

Models and Methods. We conducted a controlled experiment involving two groups of non-technical stakeholders (N-TSs) across two real-world case studies: (1) Tabiat [33], a smartphone application (and sensors) that records symptoms and keeps doctors updated on patients’ chronic obstructive pulmonary disease conditions; (2) Casper [36], a socially assistive robot designed to help people living with dementia by supporting activities of daily living at home and improving their quality of life. The first group of stakeholders, consisting of a philosopher and a psychologist, was tasked with eliciting SLEEC requirements for the Tabiat case study using RAINCOAT. The second group, consisting of a lawyer and a medical doctor, did the same, but without any tool support. For each case-study, we provided stakeholders with a detailed description of the system’s capabilities and its operating environment and conducted the following experiments:

(i) Group 1 elicited requirements without guidance for the Tabiat case study and validated them manually. We refer to this experiment as *ad hoc elicitation*. This experiment was chosen to provide a baseline.

(ii) Group 2 elicited requirements for the Casper case study following Stage I guidance of RAINCOAT and validated them manually. We refer to this experiment as *systematic-elicitation*. This was done to evaluate the impact of having a structured approach to requirements elicitation.

(iii) Group 1 and Group 2 elicited requirements for the Casper and Tabiat case studies, respectively, following the overall RAINCOAT approach (which includes guided elicitation and automated validation). We refer to these experiments as *RAINCOAT-elicitation-validation*. This experiment was aimed to measure how the systematic elicitation and the automation affect the quality of the requirements.

The elicited requirements obtained in each experiment are compared based on the size of the requirements set (including the number of events and measures) and number of well-formedness issues. To facilitate the comparison, we used LLM-San to extract relations from the manually elicited requirements (*ad hoc-elicitation* and *systematic-elicitation*), and then asked the stakeholders to filter them.

Results. Tbl. VII reports the number of requirements, events, and measures, the correctly identified semantic relations (TP), the incorrectly identified ones (FP), and the number of different WFI types for the three experiments.

The experimental results show that the requirements elicited through the systematic approach are more comprehensive and result in a larger set, compared to *ad hoc-elicitation*. In the latter case, the stakeholders were guided by the SLEEC principles and applied them to a selection of high-level system capabilities, but they did not explore all potential mappings with the complete set of system capabilities. We observed that *ad hoc-elicitation* was free from conflicts and redundancies, whereas *systematic-elicitation* had 34 WFIs, including 24 conflicts and 12 redundancies. The high number of resulting requirements in the *systematic-elicitation*, 57, rendered manual validation infeasible. In contrast, *ad hoc-elicitation* focused on a smaller set of requirements, each with distinct and independent responses.

Comparing *systematic-elicitation* with *approach-elicitation-validation*, we observe that the former yielded a significantly larger number of requirements. Since *systematic-elicitation* had to rely on manual analysis, the experimental results revealed many redundant and conflicting requirements. Moreover, we observed that when evaluating the insufficiency and restrictiveness of requirements elicited in *approach-elicitation-validation*, the stakeholders gained confidence that they had elicited ‘enough’ requirements to prevent SLEEC harms, ensuring that the system remained usable without overly restricting its main functionalities. The same stakeholders expressed

TABLE VII: RAINCOAT compared to common practice. We record: true positive (TP) and false negative (FP) in the extracted capabilities relation (rel.); the number of new issues identified (after capturing the semantic relations) for vacuous conflict (v-conf.), situational conflict (s-conf.), redundancy (redund.), restrictiveness (restrict.), and insufficiency (insuff.); and the number of iterations for each WFI type.

experiments	case studies	rules (evnt. - msr.)	relations (TP - FP)	WFI					Number of iterations					
				v-conf.	s-conf.	redund.	insuff.	restrict.	v-conf.	s-conf.	insuff.	restrict.	redund.	
<i>ad hoc-elicitation</i>	Tabiat	19 (27 - 13)	13 - 20	0	0	0	not applicable			not applicable				
<i>systematic-elicitation</i>	Casper	57 (59 - 24)	11 - 18	22	2	12	not applicable			not applicable				
RAINCOAT- <i>elicitation-validation</i>	Tabiat	28 (37 - 18)	17 - 21	0	0	0	0	0	1	3	3	0	0	
	Casper	26 (38 - 14)	23 - 22	0	0	0	0	0	0	0	1	0	1	

regret over not implementing this step during *systematic-elicitation*. For the *approach-elicitation-validation*, on average, the stakeholders converged within three iterations and resolved all instances of WFI, which required at least double the amount of time spent in *systematic-elicitation*.

We conclude that *systematic-elicitation* is crucial to elicit a comprehensive set of requirements that encompass all system capabilities. Although integrating it with automated validation (*approach-elicitation-validation*) incurred a time cost, we conclude that it was essential for preventing the elicitation of an inconsistent set of requirements. Moreover, this combined approach, implemented in RAINCOAT, effectively empowered non-technical stakeholders to produce a well-defined set of requirements without compromising the elicitation process, answering **RQ2**.

C. Threats to Validity

For the experiments answering both **RQ1** and **RQ2**, (1) the non-technical stakeholders are co-authors of this paper. We mitigated this threat by separating the authors into those who participated in developing the approach and those who evaluated it, and ensured that a complete separation between them was maintained throughout the entire lifecycle of the project. (2) Using ChatGPT for extracting semantic relations has expected risks, such as potential misinterpretation and lack of traceability. To mitigate these, we involved stakeholders in the process, asking them to review and either accept or reject the proposed relations. Although this step is time-consuming, it is still faster than a capturing the relations manually.

For the experiments answering **RQ2**, (3) the two control groups in our study included stakeholders from diverse professional backgrounds: one group consisted of a philosopher and a psychologist, and the other of a lawyer and a medical doctor. To address potential bias due to their differing expertise, we avoided making comparisons based on profession-specific requirements, focusing only on the well-formedness of the overall set of elicited requirements. (4) Our second experiment was limited to only two case studies, which could potentially narrow the scope of our conclusions. However, we mitigated this by selecting real-world case studies from different systems (i.e., a smartphone application and a robot), developed by stakeholders with diverse areas of expertise.

VI. RELATED WORK

Requirement engineering, spanning from elicitation to validation, poses significant challenges due to requirement uncertainties and potential misinterpretations by stakeholders. To address these challenges, ontology-driven techniques have

been explored, aiming to capture semantic relations among the concepts articulated in the requirements documentation [6], [13]. These approaches emphasize hierarchical relations (e.g., “is class of,” “is instance of”) between concepts [15], [37], facilitating requirements specification through relation analysis. While knowledge graphs require substantial efforts to build the entire language elements [28], [32], our approach prioritizes semantic relations, considering the domain knowledge as an input to enhance the efficacy of the elicitation process, without claiming linguistic completeness.

Recently, LLM-based techniques have emerged to streamline requirements engineering tasks, particularly in bridging the gap between natural language requirements and their specifications [3]. Investigations into the potential role of ChatGBT in the elicitation process have explored both its benefits and limitations [41]. In this paper, LLM assistance aims to extract relations among language elements of **events** and **measures**, which we term the capability relations of the system for a sound reasoning and analysis, rather than constructing a requirements specification from natural language documents [14].

VII. CONCLUSION

We have introduced a novel technique which leverages LLMs to capture common sense and bridge the gap between manually and automatically analyzing requirements. This is achieved by extracting semantic relations between the abstract representations of system capabilities in the normative requirements. These relations are then used to enrich the automated-reasoning techniques for eliciting and analyzing the consistency and coherence of normative requirements. The importance of capturing such relationships has been demonstrated by the identification of 13 new well-formedness issues (WFI) across nine existing case studies. The effectiveness and usability of our approach have been demonstrated on two real world studies with a total 64 relationships captured and 130 requirements which were elicited by stakeholders with diverse backgrounds. Our technique for extracting semantic relationships could be improved to reduce the number of false positives. In the future, we plan to explore enhancements, e.g., by refining the prompting strategies or fine-tuning the LLM to better perform the task. Our proposed approach lacks automatic support for the debugging and resolution of WFI concerns. Semi-automated generation of patches for WFIs is left for future work.

REFERENCES

- [1] Amaral, O., Azeem, M.I., Abualhaija, S., Briand, L.C.: NLP-based Automated Compliance Checking of Data Processing Agreements against GDPR. CoRR **abs/2209.09722** (2022). <https://doi.org/10.48550/arXiv.2209.09722>
- [2] Anonymous: Supplementary material for RE submission: Normative requirements operationalization with llms (2024), <https://anonymous.4open.science/r/sleecvalDef-547D/README.md>
- [3] Arora, C., Grundy, J., Abdelrazek, M.: Advancing requirements engineering through generative ai: Assessing the role of llms (2023)
- [4] Bahadır, B.N., Kasap, Z.: AutoCar Project, <https://acp317315180.wordpress.com/>
- [5] Baxter, J., Ribeiro, P., Cavalcanti, A.: Sound reasoning in tock-csp. Acta Informatica **59**(1), 125–162 (2022). <https://doi.org/10.1007/S00236-020-00394-3>, <https://doi.org/10.1007/s00236-020-00394-3>
- [6] Bencharqui, H., Haidrar, S., Anwar, A.: Ontology-based requirements specification process. <https://api.semanticscholar.org/CorpusID:266152367>
- [7] Bisk, Y., Zellers, R., Bras, R.L., Gao, J., Choi, Y.: PIQA: reasoning about physical commonsense in natural language. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020. pp. 7432–7439. AAAI Press (2020). <https://doi.org/10.1609/AAAI.V34I05.6239>, <https://doi.org/10.1609/aaai.v34i05.6239>
- [8] Calinescu, R., Alasmari, N., Gleirscher, M.: Maintaining Driver Attention in Shared-Control Autonomous Driving. In: 16th International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS). pp. 90–96. IEEE (2021)
- [9] Calinescu, R., Ashaolu, O.: Diagnostic AI System for Robot-Assisted A&E Triage (DAISY) website., <https://twitter.com/NorwichChloe/status/1679112358843613184?tr=ALk7s8wcyHztZyyHJoB5pg&s=19>, <https://tas.ac.uk/research-projects-2022-23/daisy/>
- [10] Cascella, M., Montomoli, J., Bellini, V., Bignami, E.G.: Evaluating the feasibility of chatgpt in healthcare: An analysis of multiple clinical and research scenarios. J. Medical Syst. **47**(1), 33 (2023). <https://doi.org/10.1007/S10916-023-01925-4>, <https://doi.org/10.1007/s10916-023-01925-4>
- [11] Dandy, N., Calinescu, R.: Autonomous Systems for Forest Protection (ASPEN) website., <https://tas.ac.uk/research-projects-2023-24/autonomous-systems-for-forest-protection/>
- [12] Dhingra, S., Singh, M., B. V.S., Malviya, N., Gill, S.S.: Mind meets machine: Unravelling gpt-4’s cognitive psychology. CoRR **abs/2303.11436** (2023). <https://doi.org/10.48550/ARXIV.2303.11436>, <https://doi.org/10.48550/arXiv.2303.11436>
- [13] Diamantopoulos, T.G., Symeonidis, A.L.: Enhancing requirements reusability through semantic modeling and data mining techniques. Enterprise Information Systems **12**, 960 – 981 (2018), <https://api.semanticscholar.org/CorpusID:53112505>
- [14] Fantechi, A., Gnesi, S., Passaro, L., Semini, L.: Inconsistency detection in natural language requirements using chatgpt: a preliminary evaluation. In: Proceedings of IEEE 31st International Requirements Engineering Conference (RE). pp. 335–340 (09 2023). <https://doi.org/10.1109/RE57278.2023.00045>
- [15] Farfeleder, S., Moser, T., Krall, A., Stålhane, T., Omoronyia, I., Zojer, H.: Ontology-driven guidance for requirements elicitation. In: Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., De Leenheer, P., Pan, J. (eds.) The Semantic Web: Research and Applications. pp. 212–226. Springer Berlin Heidelberg, Berlin, Heidelberg (2011)
- [16] Feng, N., Marsso, L., Getir-Yaman, S., Beverley, T., Calinescu, R., Cavalcanti, A., Chechik, M.: Towards a Formal Framework for Normative Requirements Elicitation. In: Proceedings of the 38th International Conference on Automated Software Engineering, (ASE’2023), Kirchberg, Luxembourg. IEEE (2023)
- [17] Feng, N., Marsso, L., Getir-Yaman, S., Townsend, B., Baartartogtokh, Y., Ayad, R., de Mello, V.O., Kholodetska, Y., Standen, I., Stefanakos, I., Imrie, C., Rodrigues, G., Cavalcanti, A., Calinescu, R., Chechik, M.: Analyzing and Debugging Normative Requirements via Satisfiability Checking. In: Proceedings of the 46th International Conference on Software Engineering, (ICSE 2024), Lisbon, Portugal. ACM (2024)
- [18] Feng, N., Marsso, L., Getir-Yaman, S., Townsend, B., Baartartogtokh, Y., Ayad, R., de Mello, V.O., Kholodetska, Y., Standen, I., Stefanakos, I., Imrie, C., Rodrigues, G., Cavalcanti, A., Calinescu, R., Chechik, M.: Analyzing and Debugging Normative Requirements via Satisfiability Checking. In: Proceedings of the 46th International Conference on Software Engineering, (ICSE 2024), Lisbon, Portugal. ACM (2024), <https://doi.org/10.48550/arXiv.2401.05673>
- [19] Feng, N., Marsso, L., Sabetzadeh, M., Chechik, M.: Early verification of legal compliance via bounded satisfiability checking. In: Proceedings of the 34th international conference on Computer Aided Verification (CAV’23), Paris, France. Lecture Notes in Computer Science, Springer (2023)
- [20] Getir-Yaman, S., Burholt, C., Jones, M., Calinescu, R., Cavalcanti, A.: Specification and Validation of Normative Rules for Autonomous Agents. In: Proceedings of the 26th International Conference on Fundamental Approaches to Software Engineering (FASE’2023), Paris, France. Lecture Notes in Computer Science, Springer (2023)
- [21] Getir-Yaman, S., Cavalcanti, A., Calinescu, R., Paterson, C., Ribeiro, P., Townsend, B.: Specification, validation and verification of social, legal, ethical, empathetic and cultural requirements for autonomous agents (2023), <https://arxiv.org/abs/2307.03697>
- [22] Gibson-Robinson, T., Armstrong, P.J., Boulgakov, A., Roscoe, A.W.: FDR3 - A Modern Refinement Checker for CSP. In: Proceedings of the 20th International Conference, TACAS 2014 on Tools and Algorithms for the Construction and Analysis of Systems, (TACAS’2014), Grenoble, France. Lecture Notes in Computer Science, vol. 8413, pp. 187–201. Springer (2014). https://doi.org/10.1007/978-3-642-54862-8_13
- [23] Gil, E.B., Caldas, R.D., Rodrigues, A., da Silva, G.L.G., Rodrigues, G.N., Pelliccione, P.: Body sensor network: A self-adaptive system exemplar in the healthcare domain. In: Proceedings of the 16th International Symposium on Software Engineering for Adaptive and Self-Managing Systems, (SEAMS@ICSE’2021), Madrid, Spain. pp. 224–230. IEEE (2021). <https://doi.org/10.1109/SEAMS51251.2021.00037>
- [24] Gupta, P., Khare, A., Bajpai, Y., Chakraborty, S., Gulwani, S., Kanade, A., Radhakrishna, A., Soares, G., Tiwari, A.: Grace: Language models meet code edits. In: Chandra, S., Blincoe, K., Tonella, P. (eds.) Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2023, San Francisco, CA, USA, December 3-9, 2023. pp. 1483–1495. ACM (2023). <https://doi.org/10.1145/3611643.3616253>, <https://doi.org/10.1145/3611643.3616253>
- [25] Hamilton, J., Stefanakos, I., Calinescu, R., Cámara, J.: Towards adaptive planning of assistive-care robot tasks. In: Proceedings of the Fourth International Workshop on Formal Methods for Autonomous Systems and Fourth International Workshop on Automated and Verifiable Software sYstem DEvelopment, (FMAS/ASYDE@SEFM’2022), Berlin, Germany. EPTCS, vol. 371, pp. 175–183 (2022). <https://doi.org/10.1145/4204/EPTCS.371.12>, <https://www.youtube.com/watch?v=VhfQmJe4IPc>
- [26] Haupt, C.E., Marks, M.: Ai-generated medical advice—gpt and beyond. Jama **329**(16), 1349–1350 (2023)
- [27] Hodel, D., West, J.: Response: Emergent analogical reasoning in large language models. CoRR **abs/2308.16118** (2023). <https://doi.org/10.48550/ARXIV.2308.16118>, <https://doi.org/10.48550/arXiv.2308.16118>
- [28] Hogan, A., Blomqvist, E., Cochez, M., D’amato, C., Melo, G.D., Gutierrez, C., Krrane, S., Gayo, J.E.L., Navigli, R., Neumaier, S., Ngomo, A.C.N., Polleres, A., Rashid, S.M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., Zimmermann, A.: Knowledge graphs. ACM Comput. Surv. **54**(4) (jul 2021). <https://doi.org/10.1145/3447772>, <https://doi.org/10.1145/3447772>
- [29] Jaumard, B., Simeone, B.: On the complexity of the maximum satisfiability problem for horn formulas. Inf. Process. Lett. **26**(1), 1–4 (1987). [https://doi.org/10.1016/0020-0190\(87\)90028-7](https://doi.org/10.1016/0020-0190(87)90028-7), [https://doi.org/10.1016/0020-0190\(87\)90028-7](https://doi.org/10.1016/0020-0190(87)90028-7)
- [30] Jevtic, A., Valle, A.F., Alenyà, G., Chance, G., Caleb-Solly, P., Dogramadzi, S., Torras, C.: Personalized robot assistant for support in dressing. IEEE Trans. Cogn. Dev. Syst. **11**(3), 363–374 (2019). <https://doi.org/10.1109/TCDS.2018.2817283>
- [31] Joshi, H., Sánchez, J.P.C., Gulwani, S., Le, V., Verbruggen, G., Radicek, I.: Repair is nearly generation: Multilingual program repair with llms. In: Williams, B., Chen, Y., Neville, J. (eds.) Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thir-

- teenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023. pp. 5131–5140. AAAI Press (2023). <https://doi.org/10.1609/AAAI.V37I4.25642>, <https://doi.org/10.1609/aaai.v37i4.25642>
- [32] Kejriwal, M.: Domain-specific knowledge graph construction. In: SpringerBriefs in Computer Science (2019), <https://api.semanticscholar.org/CorpusID:71147205>
- [33] Liaqat, D.: The Tabiat website., <https://www.tabiat.care/>
- [34] Lin, G., Miao, Y., Yang, X., Ou, W., Cui, L., Guo, W., Miao, C.: Commonsense knowledge adversarial dataset that challenges electra. In: 2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV). pp. 315–320. IEEE (2020)
- [35] Miller, G.A.: Nouns in wordnet: a lexical inheritance system. International journal of Lexicography **3**(4), 245–264 (1990)
- [36] Moro, C., Nejat, G., Mihailidis, A.: Learning and personalizing socially assistive robot behaviors to aid with activities of daily living. ACM Transactions on Human-Robot Interaction (THRI) **7**(2), 1–25 (2018)
- [37] Muruges, S., Jaya, A.: Construction of ontology for software requirements elicitation. Indian Journal of Science and Technology **8** (11 2015). <https://doi.org/10.17485/ijst/2015/v8i29/86271>
- [38] Nay, J.J.: Law informs code: A legal informatics approach to aligning artificial intelligence with humans. CoRR **abs/2209.13020** (2022). <https://doi.org/10.48550/ARXIV.2209.13020>, <https://doi.org/10.48550/arXiv.2209.13020>
- [39] Phung, T., Padurean, V., Cambroner, J., Gulwani, S., Kohn, T., Majumdar, R., Singla, A., Soares, G.: Generative AI for programming education: Benchmarking chatgpt, gpt-4, and human tutors. In: Fisler, K., Denny, P., Franklin, D., Hamilton, M. (eds.) Proceedings of the 2023 ACM Conference on International Computing Education Research - Volume 2, ICER 2023, Chicago, IL, USA, August 7-11, 2023. pp. 41–42. ACM (2023). <https://doi.org/10.1145/3568812.3603476>, <https://doi.org/10.1145/3568812.3603476>
- [40] Rawte, V., Sheth, A.P., Das, A.: A survey of hallucination in large foundation models. CoRR **abs/2309.05922** (2023). <https://doi.org/10.48550/ARXIV.2309.05922>, <https://doi.org/10.48550/arXiv.2309.05922>
- [41] Ronanki, K., Berger, C., Horkoff, J.: Investigating chatgpt’s potential to assist in requirements elicitation processes (2023)
- [42] Snow, R., Jurafsky, D., Ng, A.: Learning syntactic patterns for automatic hypernym discovery. Advances in neural information processing systems **17** (2004)
- [43] Stefanakos, I., Calinescu, R., Douthwaite, J.A., Aitken, J.M., Law, J.: Safety controller synthesis for a mobile manufacturing cobot. In: Proceedings of the 20th International Conference on Software Engineering and Formal Methods (SEFM’2022), Berlin, Germany. Lecture Notes in Computer Science, vol. 13550, pp. 271–287. Springer (2022). https://doi.org/10.1007/978-3-031-17108-6_17
- [44] Talmor, A., Herzig, J., Lourie, N., Berant, J.: Commonsenseqa: A question answering challenge targeting commonsense knowledge. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). pp. 4149–4158. Association for Computational Linguistics (2019). <https://doi.org/10.18653/V1/N19-1421>, <https://doi.org/10.18653/v1/n19-1421>
- [45] Townsend, B., Paterson, C., Arvind, T., Nemirovsky, G., Calinescu, R., Cavalcanti, A., Habli, I., Thomas, A.: From Pluralistic Normative Principles to Autonomous-Agent Rules. Minds and Machines pp. 1–33 (2022)
- [46] Trautmann, D., Petrova, A., Schilder, F.: Legal prompt engineering for multilingual legal judgement prediction. CoRR **abs/2212.02199** (2022). <https://doi.org/10.48550/ARXIV.2212.02199>, <https://doi.org/10.48550/arXiv.2212.02199>
- [47] Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A.T., Bi, W., Shi, F., Shi, S.: Siren’s song in the AI ocean: A survey on hallucination in large language models. CoRR **abs/2309.01219** (2023). <https://doi.org/10.48550/ARXIV.2309.01219>, <https://doi.org/10.48550/arXiv.2309.01219>