Where Norms and References Collide: Evaluating LLMs on Normative Reasoning

Mitchell Abrams[1], Kaveh Eskandari Miandoab[1], Felix Gervits[2], Vasanth Sarathy[1], Matthias Scheutz[1]

[1]Tufts University
[2]DEVCOM Army Research Laboratory
Mitchell.Abrams, Kaveh.Eskandari_Miandoab, Vasanth.Sarathy, Matthias.Scheutz{@tufts.edu}
Felix.Gervits.civ@army.mil

Abstract

Embodied agents, such as robots, will need to interact in situated environments where successful communication often depends on reasoning over social norms: shared expectations that constrain what actions are appropriate in context. A key capability in such settings is norm-based reference resolution (NBRR), where interpreting referential expressions requires inferring implicit normative expectations grounded in physical and social context. Yet it remains unclear whether Large Language Models (LLMs) can support this kind of reasoning. In this work, we introduce SNIC (*Situated Norms in Context*), a human-validated diagnostic testbed designed to probe how well state-of-the-art LLMs can extract and utilize normative principles relevant to NBRR. SNIC emphasizes physically grounded norms that arise in everyday tasks such as cleaning, tidying, and serving. Across a range of controlled evaluations, we find that even the strongest LLMs struggle to consistently identify and apply social norms—particularly when norms are implicit, underspecified, or in conflict. These findings reveal a blind spot in current LLMs and highlight a key challenge for deploying language-based systems in socially situated, embodied settings.

Code and Dataset —
https://github.com/TheSittingCat/SNIC

Introduction

Reference resolution is a critical aspect of dialogue processing that contributes to natural language understanding by linking language to real-world entities. This process is not always straightforward or sufficiently achievable through language alone. It requires interpreting language within a situated context, often consisting of social norms that are activated by objects in a scene, tasks, locations, or social roles. Prior work has shown that social norms, irrespective of language, can influence the interpretation of a referent (Abrams and Scheutz 2022).

Consider an example where a human asks an assistant in a communal kitchen: Human: *"Can you pass me a mug?"* Assistant: *"Okay."* There are three mugs nearby—one clean and two with leftover coffee. Without additional linguistic cues, the request may seem ambiguous. However, given the

context, the intended referent is likely the clean mug because, *normatively*, people avoid drinking from used mugs that are not their own or dirty. The underlying norm applied to this situation is: *you should use clean items when drinking or serving beverages*. If the context instead involved cleaning up after an event, then the request might be understood as referring to the dirty mugs. Without normative reasoning, this interpretation could be lost.

Embodied agents or AI systems must incorporate social norms into reference resolution as they operate in social environments. In interactive situations, *social norms* serve as powerful constraints because they consist of shared expectations that people rely on to coordinate behavior. Norms function similarly to language in that they provide a type of social grammar (Bicchieri 2005), guiding interlocutors in resolving referential ambiguity. However, normative reasoning for reference resolution is a challenging task; norms are highly contextual and range from abstract (e.g., *show respect*) to very specific and physically grounded (e.g., *you should pick up your glass by the stairs so nobody knocks it over*). Norms also vary in how strongly they should be followed or violated[1], which is particularly relevant in norm conflicts where some norms may need to be violated to satisfy stronger norms (Malle, Bello, and Scheutz 2019).

For reference resolution, norms must be identified from both language and context, then reasoned with to resolve a referent. Large Language Models (LLMs) provide a promising solution for norm identification and reasoning since they implicitly encode world knowledge and perform well on question-answering tasks (Brown et al. 2020). However, norms may be difficult to learn through exposure to large amounts of text, as they are often implicit rather than explicitly stated. It is an open question whether LLMs can perform reference resolution that requires normative reasoning.

To investigate this challenge, we introduce SNIC, a human-validated diagnostic testbed designed to represent the prescriptive and prohibitive social norms that guide referential interpretation in everyday contexts. Unlike prior norm datasets that focus on abstract moral judgments (Forbes et al. 2020), SNIC emphasizes physically grounded norms embedded in real-world tasks that humans and agents would engage in (e.g., cleaning, serving, tidying). Rather

---

[1]referred to as *deontic force*

than proposing SNIC as a definitive benchmark, we use it as a controlled environment to probe how LLMs handle normative ambiguity, prioritize between conflicting norms, and respond to implicit expectations when resolving ambiguous referential expressions. Across a series of evaluations, we find that models exhibit inconsistencies; not only across different scenarios but even within near-identical scenes under slightly varied conditions. These findings expose a critical but unexplored weakness in contemporary language models: their ability to consistently reason about physically situated social norms.

Background

Social norms function as a form of social grammar, shaping expectations for behavior in different contexts (Bicchieri 2005). This is because social norms consist of both *empirical expectations* (beliefs about what others do) and *normative expectations* (beliefs about what others expect us to do) (Malle, Bello, and Scheutz 2019). Their power to guide behavior has been attractive to the AI community.

Normative reasoning has been studied in AI and the NMAS (normative multi-agent systems) community through formal representations that encode norms using *deontic logic* and contextual activation conditions (Malle, Bello, and Scheutz 2019; Sarathy et al. 2017; Malle et al. 2020). These studies have explored how norms are triggered by environmental and linguistic cues, but their application to reference resolution remains unexplored.

To our knowledge, no existing dataset explicitly evaluates *norm-based reference resolution* (NBRR). There has been NLP work in anaphora and coreference resolution that requires external commonsense knowledge for interpretation: reference resolution tasks, such as the *Winograd Schema Challenge* (Levesque, Davis, and Morgenstern 2012), *Winogrande* (Sakaguchi et al. 2021) (a reference resolution task that requires common-sense knowledge to resolve ambiguous references), and *CommonsenseQA* (Talmor et al. 2018), test common-sense reasoning but do not isolate normative constraints. We follow this tradition and extend the idea that reference and co-reference resolution require not only commonsense knowledge, but even normative knowledge. We also treat text-based vignettes as a reasonable proxy for physical scenarios, as in prior work like TextWorld (Côté et al. 2019), though our focus remains on evaluating norms in text form.

In social norm modeling, NLP systems such as *Delphi* and the *Neural Norm Transformer* have been trained on datasets like *Social Chemistry 101* (Forbes et al. 2020). However, these models primarily capture norms related to moral judgments and social etiquette, lacking the physical and situational grounding needed for norm-based reference resolution. This gap motivates the development of new testbeds and evaluation methods tailored to NBRR This dataset, for instance, consists of situations from online forums (e.g., *"Asking my boyfriend to stop being friends with his ex"*) and associated moral rules of thumb (e.g., *"It's okay to ask your significant other to stop doing something you're uncomfortable with"*) This gap motivates the development of new benchmarks and evaluation methods tailored to NBRR.

LLMs offer a potential solution for addressing the challenges of situated and norm-based reference resolution as they can capture higher-level contextual information and world knowledge and perform well on question-answering tasks and common-sense reasoning (Brown et al. 2020). He et al. (2024) evaluated LLMs on detecting norm violations in household task sequences, using textual scenarios with human-annotated ground truth. Their task focused on post-hoc classification of whether an agent's actions violated a given norm, with results ranging from 62.5% to 100% accuracy across generic and role-based norms using ChatGPT-4. This suggests that LLMs can vary in the types of norms they can detect. Our work, in contrast, addresses a distinct challenge: resolving ambiguous referring expressions where the correct interpretation hinges on proactively applying implicit or conflicting social norms.

There is still other work that highlights LLMs and general social reasoning. Bian et al. (2023) shows LLMs struggle more with social common sense reasoning with the SocialIQA dataset (Sap et al. 2019) and social factors of language outside of content (Hovy and Yang 2021), in addition to tasks involving implicit social norms and physical grounding (Bian et al. 2023; Sap et al. 2019). Other work has looked the LLM evaluations of moral dilemmas which are related to social norms (Sachdeva and van Nuenen 2025).

The integration of LLMs into robotics and embodied AI suggests potential applications for norm-based reference resolution. In robot dialogue systems, LLMs have been used for *object disambiguation* (Jiang, Xu, and Hsu 2024), *command disambiguation* (Park et al. 2023), and *household organization tasks* (Wu et al. 2023). However, even in these applications, LLMs struggle with handling norm-relevant distinctions. For example, Wu et al. (2023) found that when generating user preference rules for tidying, LLMs occasionally overgeneralized norms, leading to incorrect object groupings (e.g., treating all drawers as equivalent despite cultural differences in organization). These limitations highlight why NBRR remains an open problem: while LLMs can encode broad world knowledge, it is unclear whether they reliably identify and apply context-specific norms for NLP tasks like reference resolution.

Existing datasets in reference resolution and social norm modeling fail to integrate the physical and situational constraints required for resolving referents in a normative context. Given these gaps, our work introduces **SNIC**, a testbed designed to evaluate NBRR, and investigates how well LLMs can extract and apply norms in one-shot reference resolution tasks.

Dataset Generation and Augmentation Pipeline

Creating large-scale datasets for norm-based reference resolution (NBRR) is time-intensive and costly. To ensure data quality and alignment with social norms, we first conducted a human validation study on an initial handcrafted set of 120 examples (filtered to n=51), which served as the foundation for a procedurally augmented corpus of 9,000 examples. This section describes our dataset creation pipeline, includ-

ing human validation, reference resolution annotation, and augmentation strategies (see Figure 2 for an overview).

To ensure that reference resolution depends on social norms, each scene was designed with two core requirements: (1) the referring expression must be ambiguous, and (2) resolving the referent must rely on implicit or explicit social normative knowledge. For example, in a scene with a clean and dirty plate, the ambiguous utterance *"pick up the plate"* requires knowledge of the norm *"one should not use dirty kitchen items"* to disambiguate the referent.

Ambiguity is a deliberate feature of SNIC. The goal of each instance is not to enforce a single linguistically correct answer, but to test whether models consistently select the norm-guided referent even when multiple interpretations remain pragmatically possible. Our analysis of partial-match selections in the human study demonstrates that norms still exert a strong influence under ambiguity.

To validate that referent selection is guided by social norms, we conducted an online study with 210 participants recruited via Prolific. Participant ages ranged from 21 to 69, approximately balanced by gender. Data was anonymized, and participants could withdraw at any time. Compensation was $1.00 for a 5-minute study that was IRB approved. Figure 1 showcases a sample vignette from our human-subject experiment. In the main experiment, each participant was assigned six scenes in a randomized order (so they encountered different norms). Participants first read an initial prompt (*"You will be presented with a series of textual vignettes that describe a human-robot interaction. You will then be asked questions related to objects in the scene. In your answer, check the option that applies (you may also select multiple options if applicable"*). Then they were asked to read the context, select the intended referent (*"What object is the speaker referring to¿ "*), and provide a free-response justification (*"Please give reasoning for your response"*). Since selecting the referent inherently relied on norm understanding, referent choice serves as an *implicit validation mechanism*.

Step 1: Seed Dataset Elicitation (n=120)

To evaluate whether our hypothesized norm-guided referents align with human interpretation, we conducted a reference resolution study using 120 scenario-based questions grouped into nine distinct norm reasoning categories. Each scenario presents a textual vignette with a setting, a task (e.g., cleaning, cooking), and 5–7 potential referent objects with relevant properties. The scenario also includes a linguistically ambiguous referring expression embedded in an imperative utterance (e.g., *"hand me the book"*), where the correct referent is determined by an underlying social norm.

The dataset is constructed to emphasize physically grounded social norms, such as: *Do not touch something that does not belong to you*, *Do not use decorative items*, *You should clean kitchen items that are dirty*, and *Avoid dangerous scenarios*.[2] These norms are instantiated across multi-

---
[2]Following Dignum and Kuiper (1997), some norms prescribe an action to rectify another norm violation (e.g., removing broken glass on the floor).

ple settings, including *restaurant*, *kitchen*, and *library*, with variations in object properties and candidate referents.

Below are the social norms and norm conflict scenarios represented in the dataset. We follow Olson, Salas-Damian, and Forbus (2024) in defining our norm conflicts as *direct norm conflict* as two norms that apply within a context that have the same behavior (e.g. cleaning action):

Norms
Norm 1: *In a serving task, you should serve with kitchen items that are clean*
Norm 2: *In a cleaning task, you should clean items that are dirty*
Norm 3: *In a cooking task, you should cook with tools that are clean*
Norm 4: *In a serving task you should not use items that are decorative*
Norm 5: *In a cleaning task, you should not clean items that are currently being used by someone*
Norm 6: *In a cleaning task you should prioritize cleaning up any hazards (e.g. a broken plate on the ground)*
Norm 7: *In general, you should prioritize cleaning up any hazards*
Norm Conflict 1: *In a tidying task, you should prioritize cleaning up hazards over dirty items*
Norm Conflict 2: *In a tidying task in a library, you prioritize removing dirty items from the floor over books and other objects*

Step 2: Validation Analysis and Seed Set Filtering (n=51)

To assess how often our hypothesized norm-guided referent matched participant selections, we computed three levels of agreement per question: *full match* (the norm-guided referent was selected exclusively), *partial match* (the norm-guided referent was selected alongside others), and *no match* (it was not selected at all). We summarize the results across norm groups in Table 1. We measured agreement across 9–10 annotators per question.

Because participants could select multiple referents per question, we adapted Fleiss' κ—a standard inter-rater agreement metric—by binarizing responses: a 1 if the norm-conforming referent was selected, and 0 otherwise. The κ values in Table 1 reflect inter-annotator agreement on this binary decision, approximating whether participants converged on norm-based reasoning. Partial matches notably provide evidence that participants often included the norm-guided referent even when ambiguity led them to select others as well.

We only used scenes where the norm-guided referent received the most or tied-most votes (i.e., plurality winner) in the human study as seeds for augmentation, ensuring that the expanded dataset inherits empirically grounded referential interpretations.

Table 1 presents agreement statistics for the filtered set of 51 questions used in our core dataset, grouped by normative category. Match rates (% Match) range from 31.1% to 64.7%

Norm	Questions	Total Votes	% Match	% Partial	% No Match	Kappa*
Norm 1	16	145	58.6	33.1	8.3	0.215
Norm 2	10	88	48.9	34.1	17.0	0.166
Norm 3	6	57	57.9	24.6	17.5	0.073
Norm 4	2	17	52.9	29.4	17.6	-0.113
Norm 5	2	17	64.7	23.5	11.8	-0.112
Norm 6	4	35	37.1	42.9	20.0	0.033
Norm 7	3	29	41.4	34.5	24.1	-0.074
NC 1	3	27	37.0	44.4	18.5	-0.058
NC 2	5	45	31.1	40.0	28.9	-0.013

Table 1: Participant agreement per norm group using per-question hypothesized referents. These are "Plurality Winner" questions in which the normative option received the most votes by annotators. Fleiss' κ reflects inter-annotator agreement on exact referent strings. Partial matches include all responses that contain the norm-guided referent alongside others. Agreement metrics are reported at the participant response level.

across norm groups, with many norms also showing substantial partial matches, where participants selected the norm-guided referent alongside other options. For instance, Norm 1 (*In a serving task, serve with clean kitchen items*) yielded 58.6% match and 33.1% partial, indicating strong—but pluralistically expressed—alignment with the hypothesized referent. Norms such as Norm 2 and Norm 3 also showed high match and partial agreement, with κ values suggesting modest but consistent convergence across annotators.

Other norms, such as Norms 6 and 7, exhibited lower κ and match rates, reflecting greater variability in how participants interpreted the norms' implications. These results align with our theoretical expectation that not all norms are equally salient or widely shared—some, such as cleanliness in cooking, appear robust and universally interpretable, while others (e.g., hazard prioritization or ownership constraints) are more context-sensitive.

Overall, these results validate the hypothesis that social norms influence referential interpretation, while also revealing variation in how strongly different norms are encoded and applied by human participants.

The robot is assisting in a **cooking task** in a kitchen.
The human instructs the robot: "***hand me the spatula***"

The robot detects 7 objects:

Object Number	Object Description
1	a clean spoon on the counter that is communal
2	a dirty spatula on the counter that is communal
3	a clean spatula on the counter that belongs to the speaker
4	a clean fork on the counter that is communal
5	a dirty spatula in the sink that belongs to the speaker
6	a clean towel on the counter that is communal
7	a clean bowl in the cabinet that is communal

Figure 1: Sample text-vignette from human-subject experiment validating our dataset. This question is a reference task with the social norm *In a cooking task, you should cook with tools that are clean*.

The free text justifications provided some evidence of more explicit norm justification:
"It is a serving task and you should always serve with a clean utensil."
"Tidying up library books is my assumption so I chose the dictionary. I also assume the students are near their notebooks so we shouldn't bother them."
"If this is a cleaning task, the book on the ground would be what the robot would have to clean."
"If it's a cleaning task, then the person must mean to give them the dirty spoon so they can clean it."

Step 3: Procedural Augmentation of 9,000 Cases

To scale the dataset while maintaining norm consistency, we employed a controlled procedural data augmentation strategy inspired by Thierauf, Abrams, and Scheutz (2024). This process systematically introduced variations in object types, object properties, and settings while ensuring that referent choices remained norm-governed based on the smaller human-validated subset. A sample template rule generated scenes based on the validated norm represented in Figure 1 from our original study. F-strings (f) in Python allow variables and expressions to be embedded directly within a string:

```python
prompt = (
  f"{name_pair[0]} and {name_pair[1]} are in the {location}. "
  f"{name_pair[0]} is cooking. "
  f"{name_pair[1]} is next to {', '.join(objects)}. "
  f"{name_pair[0]} asks {name_pair[1]} to hand them {
      referring_expression}."
)
```

This rule, for instance, creates a textual representation that simply states the main actors, the setting, the task, the objects with their relevant properties, and the referring expression:

*Jordan and Taylor are in the kitchen. Jordan is cooking. Taylor is next to calculator, clean mug, dirty mug, pen, dirty mug, notebook. Jordan asks Taylor to **hand them the mug***

Each scene has a corresponding formal representation so that input representation could be compared in the evaluation (natural language vs a formally precise language). This

symbolic representation aims to eliminate any ambiguities about any elements of the scene. Below is an example representation of the scene above.

```
1   setting(kitchen).
2   task(cooking).
3   speaker(jordan, human).
4   listener(taylor, human).
5   object(obj1).
6   property(obj1, calculator).
7   object(obj2).
8   property(obj2, clean).
9   property(obj2, mug).
10  object(obj3).
11  property(obj3, dirty).
12  property(obj3, mug).
13  object(obj4).
14  property(obj4, pen).
15  object(obj5).
16  property(obj5, dirty).
17  property(obj5, mug).
18  object(obj6).
19  property(obj6, notebook).
20  referring_expression(the mug).
```

Unlike large-scale reference resolution datasets such as WinoGrande (Sakaguchi et al. 2021), which rely on crowd-sourcing, our method follows structured rule-based augmentation techniques commonly used in NLP (Wei and Zou 2019). While the expanded dataset (n=9,000) was not independently validated, it adheres to patterns established in the human-validated subset, ensuring reliability. By combining human validation with procedural augmentation, we balance scalability with interpretability, making the dataset well-suited for evaluating norm-based reference resolution.

3.1: Ground Truth and Quality Assurance Regarding ground truth and scale, all 9,000 augmented examples are generated from a set of 51 human-validated seeds where the normative referent was the plurality choice. As such, the ground truth for these 9,000 instances is established by construction. The augmentation procedure varies only non-essential elements (e.g., substituting object types while preserving properties such as "dirty" or "clean"), ensuring that each derived scene instantiates the same core normative structure as its validated seed. While additional downstream validation would be beneficial, the rule-based augmentation approach provides a reliable mechanism for producing large-scale normative contexts. Our focus on static, one-shot reasoning is intentional: it isolates what models can infer from context alone, without the aid of clarification-seeking strategies and multi-turn dialogue. Finally, our emphasis on everyday norms that are physically grounded and our selection of contemporary LLMs (Llama-3 (Touvron et al. 2023), Phi-3 and Phi-4 (Abdin et al. 2024), and GPT-4o-mini (Achiam et al. 2023), 2024–2025 releases) reflects our goal of building a controlled and interpretable diagnostic testbed for assessing current normative reasoning capabilities.

Category	Distribution (%)
Object Count	
5 objects	14.13
6 objects	42.79
7 objects	28.4
8 objects	14.69
Task Distribution (for 8,000 scenes)	
Tidying	12.50
Cleaning	50.00
Serving	25.00
Cooking	12.50
Setting Distribution	
Living room	15.83
Restaurant	15.37
Dining room	14.84
Garden	14.64
Banquet hall	14.12
Kitchen	14.08
Library	11.11
Distinct Object & Properties Count	
Distinct Objects	46
Distinct Settings	7
Explicit Properties	7
Referring Expressions	
hand me/them the X	5,000
pick it up	4,000

Table 2: Dataset summary: % object type distribution, task distribution, and setting distribution across dataset.

Dataset

Our final dataset artifact (SNIC) contains 9,000 questions, divided into nine categories of 1,000 questions each. Each category represents a social norm or norm conflict embedded in the scene. We designed these larger norm categories as opposed to a dataset with greater norm diversity (i.e., a different norm for every question) because it would make it more challenging to systematically compare the effects of a specific norm.

Table 2 shows the % distribution of the number of candidate objects across the whole corpus (maintaining the same object range as the human-validated dataset), and the % distribution of tasks and settings. The expanded dataset consists of more diverse settings (e.g., garden, kitchen, living room) and variety of object types and properties, ensuring broader generalization.

Evaluations

We first sample a list of models such that a diverse set of model families and sizes is represented. The models under test include Granite-3.1 2B and 8B (Mishra et al. 2024), Phi-3 3.8B and 14B, Phi-4 3.8B (Abdin et al. 2024), Llama-3 3B (3.2 variant) and 8B (3.1 variant) (Touvron et al. 2023), Qwen-2.5 3B, 7B, and 14B (Qwen et al. 2025), and GPT-4o-mini (Achiam et al. 2023).

With the exception of GPT-4 model versions, all models are tested via the invocation of Ollama using DSPY (Khattab
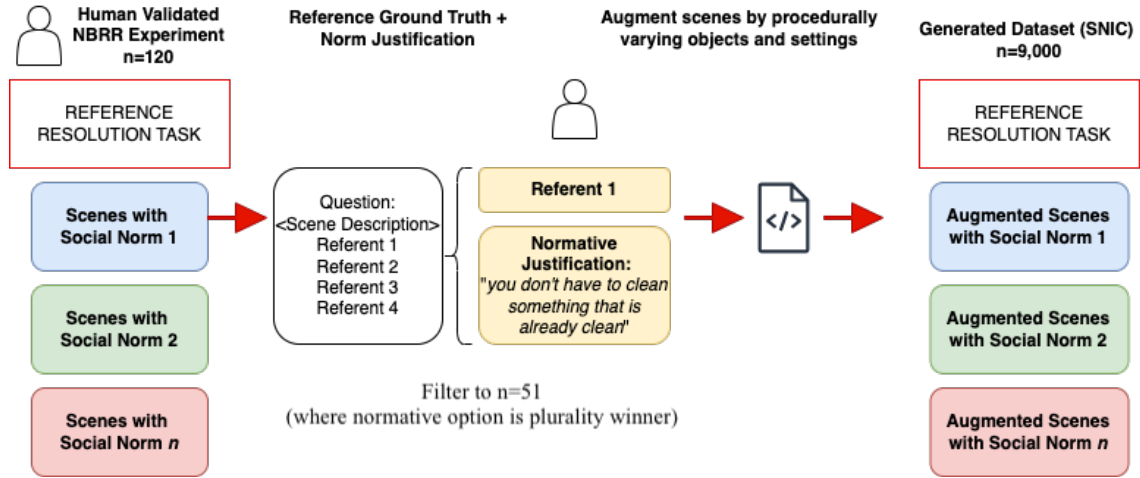
DATASET AUGMENTATION PIPELINE

Figure 2: Data augmentation pipeline to create our final dataset (SNIC) (n=9,000) for LLM evaluation.

et al. 2024) in their 4-bit quantized form and with a temperature of 0 and a seed of 0 on a single A6000 GPU with 48GB VRAM. GPT-4o variants are evaluated in their original form as served by OpenAI.

We evaluate each model under three general settings. First, to assess the normative knowledge of LLMs, we compute the model accuracy given the entirety of each instance (scene description + Prolog (FOL) code). The purpose of the Prolog formal specification is to test input types and to make the environment very precise (e.g. assigning ids to the candidate referent objects so it is clear which one should be referenced even if there are multiple objects of the same type e.g. cup). This was ultimately aimed at removing ambiguities in the scene. Second, in order to understand the effects of formalization on a model's performance, we eliminate the Prolog formalization from each instance, expecting the model to adhere to social norms through scene description. Finally, we explicitly augment the model's context with the expected social norms to follow (listing all social norms), testing the hypothesis that the lack of social norm knowledge in LLMs is a contributing factor to their performance. The Appendix showcases the model instructions used for each setting.

Table 3 showcases the performance of different models under our evaluation settings. We find that generally, when a model is provided with the scene description as well as the corresponding formalization, it is not able to consistently infer the correct social norm, with GOT-4.1 performing better than other models at 59.62% with an average of 44.08% across all models.

Somewhat surprisingly, Prolog formalization of the problem does not seem to significantly affect the model performance in our task, with the average performance being 44.22% across models, and the top performing model (GPT4.1) having an accuracy of 56.98% in the cases where only the scene description is provided to the Large Language Model. This observation might be due to the fact that LLMs

do not have a strong understanding of Prolog as a descriptive language, or stem from their lack of socially normative reasoning knowledge such that additional formalization provides no further hints for the model to act upon. Two notable exceptions to this rule are Llama-3 3B, which suffers a significant performance degradation when supplied with a Prolog formalization, and Granite-3 2B, which enjoys a performance increase in such a case. We attribute these observed variations in performance in these models toward their general understanding of formalization as well as their general sensitivity to changes in the input prompt (Shi et al. 2023; Eskandari Miandoab and Sarathy 2024). Further testing is required against other forms of formalization to fully understand the effects of formalizing language on LLM behavior.

To further probe the LLM knowledge of social norms, we evaluate each model in the setting that it is provided with the norms to follow in its context in addition to the scene description and the Prolog formalization. We find that providing the list of social norms to follow and asking it to adhere to these rules significantly increases the LLM performance, with the average accuracy rising to 70.51%, with Llama-3 3B being the only model that does not see a significant performance increase. This observation bolsters our hypothesis that current LLMs do not inherently have a strong knowledge of implicit social norms to follow, although they are able to follow them when provided, raising the need for further alignment with respect to these norms in their training stage in the age of agentic models that can directly interact with, and affect individuals (Kapoor et al. 2024). A fine-grained analysis is provided in the Appendix.

Furthermore, we observe an interesting correlation between a number of social norms as defined in our datasets. Figure 3 showcases the Spearman correlation between the social norms with respect to model performance. We see that several categories have a strong positive or negative correlation with other social norms. While the positive correlations
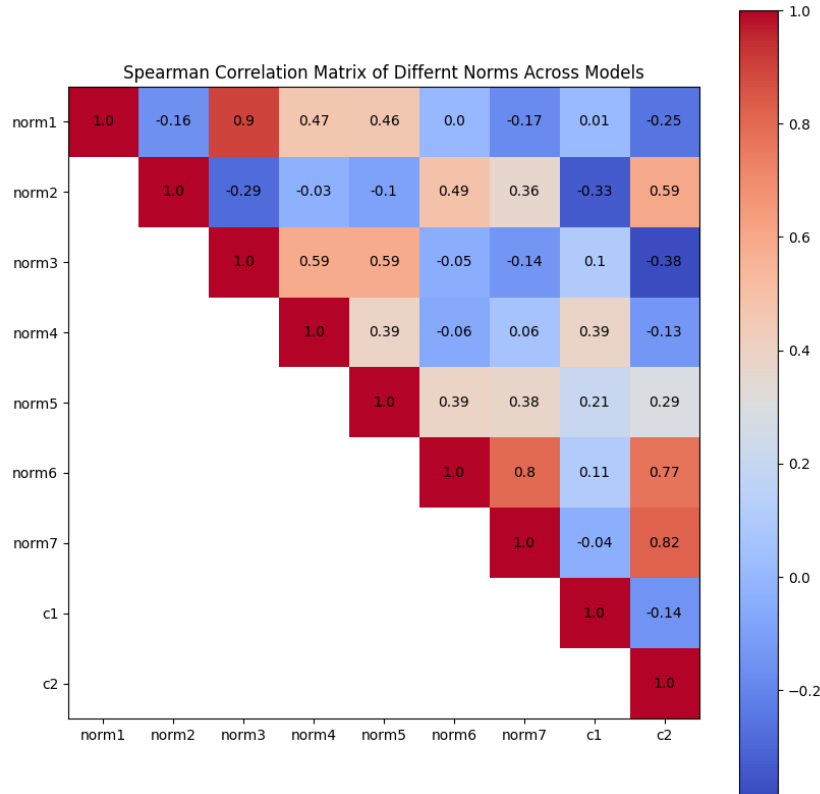
Figure 3: Spearman Correlation of Social Norms

can be attributed to the structural similarity of some norms (Norm 1 and Norm 3 for instance as defined previously), the strong negative correlation between a number of norms (Norm 3 and Norm Conflict 1 for example), is particularly interesting, as it shows clear precedence in the inference procedure of the models, where they prioritize a certain task, even in conditions that impose a rearrangement in the prioritization (safety over tidiness). This phenomenon is likely due to the lack of a strong understanding of social norms by the models, as well as a weaker emphasis on the correct ranking of social and ethical norms during the alignment process.

Lastly, we report model accuracy on the original SNIC subset of 51 human-validated items without additional context (i.e., no norm list or formal representation) in Table 4. The overall performance is slightly higher than on the expanded dataset and supports the value of dataset expansion: the larger dataset introduces greater variation and challenging distractors, making it harder for models to consistently identify the correct norm. These results suggest that scaling the dataset was not only useful, but necessary for evaluating normative reasoning in more complex contexts.

Discussion & Conclusion

In this work, we introduced Situated Norms In Context (SNIC), a dataset of 9,000 instances augmented from 51 human curated scenarios. SNIC focuses on a specific set of physically grounded social norms that rely on understanding of human social norms to solve, and aims for controlled evaluation and quantification of AI systems in the context of social norm understanding in everyday tasks.

Our results show that LLMs struggle to perform consistently well on our testbed unless explicitly provided with social norms. This suggests that detecting and reasoning with physically grounded social norms in text for reference resolution remains a challenging capability. One likely explanation is that many such norms are underrepresented or implicit in the textual data that LLMs are trained on. While LLMs excel at commonsense reasoning (Brown et al. 2020), likely due to exposure to frequently stated facts and rou-

Model	Description Only	Description + FOL	Description + FOL + Norms
Granite-3 2B	27.52	46.65	60.66
Granite-3 8B	47.76	46.06	69.98
Phi-3 3.8B	40.6	37.56	70.86
Phi-3 14B	53.61	51.54	77.12
Phi-4 3.8B	53.95	38.73	54.36
Llama-3 3B	33.71	12.01	26.6
Llama-3 8B	41.53	46.56	75.53
Qwen-2.5 3B	43.38	45.17	69.92
Qwen-2.5 7B	54.06	43.35	67.66
Qwen-2.5 14B	44.78	45.05	80.35
GPT-4o-mini	42.58	51.41	93.57
GPT-4.1	**56.98**	**59.62**	**99.6**

Table 3: Model performance under different testing methods. The formal specification we add to the input is a first-order logic representation (FOL).

Model Name	Accuracy (%)
Granite 3.1 2B	52.49
Granite 3.1 8B	39.21
Phi-3 3.8B	52.94
Phi-4 3.8B	52.94
Phi-3 14B	58.84
Llama 3 3B	66.66
Llama 3 8B	54.90
Qwen 2.5 3B	47.05
Qwen 2.5 7B	**68.67**
Qwen 2.5 14B	45.09
GPT-4o Mini	45.09

Table 4: Model accuracy on the original SNIC subset (51 human-validated items). This evaluation reflects reference resolution performance without norm or FOL augmentation.

tines, social norms differ from both standard conventions and common-sense expectations. Certain prohibitions are rarely stated explicitly (e.g., *do not use a stranger's toothbrush*), while general prescriptive norms (e.g., *show respect*) are too abstract to encode consistently. Norms also carry implicit expectations that are culturally shared, making them difficult to learn from text alone.

LLMs lack of normative understanding in physically grounded contexts may be also attributed to the lack of sufficient texts or corpora that describe references that happen in norm-specific, everyday contexts. Because there is no large-scale corpus of physically situated, norm-sensitive references, this gap is not easily remedied by additional training alone. Our dataset is particularly challenging as each example reflects a specific, physically instantiated norm rather than an abstract moral or social rule. Unlike prior datasets such as Social Chemistry 101 (Forbes et al. 2020), which focus on generalized social rules, SNIC emphasizes physically situated contexts. While this does not directly evaluate real-world deployment, it suggests limitations in LLMs' normative reasoning abilities that could impact situated applications requiring interpretation of implicit social expectations.

Improving LLMs' ability to reason about such norms could benefit downstream applications in assistive robotics and autonomous agents by making models more socially aware and context-sensitive. It could also strengthen human-agent language understanding by enabling models to adhere to implicit societal expectations and resolve ambiguities in more human-like ways (Hovy and Yang 2021).

Limitations

Some limitations suggest directions for future work. First, our testbed is text-based and does not evaluate multimodal reasoning with perceptual cues or physical affordances, which are critical for embodied agents. Second, while we evaluate pretrained models, we do not explore alternative training paradigms such as reinforcement learning, human feedback alignment, or explicit norm supervision, which might improve normative sensitivity. The goal of this current work, however, is to assess how well existing LLMs handle norm-based reference resolution, rather than to develop new training methods. Third, although SNIC includes human validation, we do not employ structured norm elicitation or cross-participant agreement protocols to assess norm consensus, limiting the granularity of our behavioral grounding. Future work should explore more rigorous validation methods to ensure a stronger consensus on the underlying norms guiding reference resolution. Finally, given the rapid progress in the field of Large Language Models, it is difficult to make a general statement regarding the norm reference capabilities of all state-of-the-art LLMs. Potentially including social norm understanding as a part of the continuous evaluation of such models can serve as a valuable resource in assessing their usability in social and collaborative settings.

Acknowledgments

References

Abdin, M.; Aneja, J.; Behl, H.; Bubeck, S.; Eldan, R.; Gunasekar, S.; Harrison, M.; Hewett, R. J.; Javaheripi, M.; Kauffmann, P.; Lee, J. R.; Lee, Y. T.; Li, Y.; Liu, W.; Mendes, C. C. T.; Nguyen, A.; Price, E.; de Rosa, G.; Saarikivi, O.; Salim, A.; Shah, S.; Wang, X.; Ward, R.; Wu, Y.; Yu, D.; Zhang, C.; and Zhang, Y. 2024. Phi-4 Technical Report. arXiv:2412.08905.

Abrams, M.; and Scheutz, M. 2022. Social Norms Guide Reference Resolution. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1–11.

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bian, N.; Han, X.; Sun, L.; Lin, H.; Lu, Y.; He, B.; Jiang, S.; and Dong, B. 2023. Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. *arXiv preprint arXiv:2303.16421*.

Bicchieri, C. 2005. *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Côté, M.-A.; Kádár, A.; Yuan, X.; Kybartas, B.; Barnes, T.; Fine, E.; Moore, J.; Hausknecht, M.; El Asri, L.; Adada, M.; et al. 2019. Textworld: A learning environment for text-based games. In *Computer Games: 7th Workshop, CGW 2018, Held in Conjunction with the 27th International Conference on Artificial Intelligence, IJCAI 2018, Stockholm, Sweden, July 13, 2018, Revised Selected Papers 7*, 41–75. Springer.

Dignum, F.; and Kuiper, R. 1997. Combining dynamic deontic logic and temporal logic for the specification of deadlines. In *Proceedings of the Thirtieth Hawaii International Conference on System Sciences*, volume 5, 336–346. IEEE.

Eskandari Miandoab, K.; and Sarathy, V. 2024. "Let's Argue Both Sides": Argument Generation Can Force Small Models to Utilize Previously Inaccessible Reasoning Capabilities. In Kumar, S.; Balachandran, V.; Park, C. Y.; Shi, W.; Hayati, S. A.; Tsvetkov, Y.; Smith, N.; Hajishirzi, H.; Kang, D.; and Jurgens, D., eds., *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, 269–283. Miami, Florida, USA: Association for Computational Linguistics.

Forbes, M.; Hwang, J. D.; Shwartz, V.; Sap, M.; and Choi, Y. 2020. Social chemistry 101: Learning to reason about social and moral norms. *arXiv preprint arXiv:2011.00620*.

He, S.; Ranathunga, S.; Cranefield, S.; and Savarimuthu, B. T. R. 2024. Norm violation detection in multi-agent systems using large language models: A pilot study. In *International Workshop on Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems*, 146–159. Springer.

Hovy, D.; and Yang, D. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, 588–602.

Jiang, C.; Xu, Y.; and Hsu, D. 2024. LLMs for Robotic Object Disambiguation. *arXiv preprint arXiv:2401.03388*.

Kapoor, S.; Stroebl, B.; Siegel, Z. S.; Nadgir, N.; and Narayanan, A. 2024. AI Agents That Matter. arXiv:2407.01502.

Khattab, O.; Singhvi, A.; Maheshwari, P.; Zhang, Z.; Santhanam, K.; Vardhamanan, S.; Haq, S.; Sharma, A.; Joshi, T. T.; Moazam, H.; Miller, H.; Zaharia, M.; and Potts, C. 2024. DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines.

Levesque, H.; Davis, E.; and Morgenstern, L. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Malle, B. F.; Bello, P.; and Scheutz, M. 2019. Requirements for an artificial agent with norm competence. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 21–27.

Malle, B. F.; Rosen, E.; Chi, V. B.; Berg, M.; and Haas, P. 2020. A General Methodology for Teaching Norms to Social Robots. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 1395–1402. IEEE.

Mishra, M.; Stallone, M.; Zhang, G.; Shen, Y.; Prasad, A.; Soria, A. M.; Merler, M.; Selvam, P.; Surendran, S.; Singh, S.; Sethi, M.; Dang, X.-H.; Li, P.; Wu, K.-L.; Zawad, S.; Coleman, A.; White, M.; Lewis, M.; Pavuluri, R.; Koyfman, Y.; Lublinsky, B.; de Bayser, M.; Abdelaziz, I.; Basu, K.; Agarwal, M.; Zhou, Y.; Johnson, C.; Goyal, A.; Patel, H.; Shah, Y.; Zerfos, P.; Ludwig, H.; Munawar, A.; Crouse, M.; Kapanipathi, P.; Salaria, S.; Calio, B.; Wen, S.; Seelam, S.; Belgodere, B.; Fonseca, C.; Singhee, A.; Desai, N.; Cox, D. D.; Puri, R.; and Panda, R. 2024. Granite Code Models: A Family of Open Foundation Models for Code Intelligence. arXiv:2405.04324.

Olson, T.; Salas-Damian, R.; and Forbus, K. D. 2024. A Defeasible Deontic Calculus for Resolving Norm Conflicts. *arXiv preprint arXiv:2407.04869*.

Park, J.; Lim, S.; Lee, J.; Park, S.; Chang, M.; Yu, Y.; and Choi, S. 2023. CLARA: classifying and disambiguating user commands for reliable interactive robotic agents. *IEEE Robotics and Automation Letters*.

Qwen; :; Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Tang, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2025. Qwen2.5 Technical Report. arXiv:2412.15115.

Sachdeva, P.; and van Nuenen, T. 2025. Normative evaluation of large language models with everyday moral dilemmas. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 690–709.

Sakaguchi, K.; Bras, R. L.; Bhagavatula, C.; and Choi, Y. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9): 99–106.

Sap, M.; Rashkin, H.; Chen, D.; LeBras, R.; and Choi, Y. 2019. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Sarathy, V.; Scheutz, M.; Kenett, Y. N.; Allaham, M.; Austerweil, J. L.; and Malle, B. F. 2017. Mental Representations and Computational Modeling of Context-Specific Human Norm Systems. In *CogSci*, volume 1, 1–1.

Shi, F.; Chen, X.; Misra, K.; Scales, N.; Dohan, D.; Chi, E.; Schärli, N.; and Zhou, D. 2023. Large Language Models Can Be Easily Distracted by Irrelevant Context. arXiv:2302.00093.

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Thierauf, C.; Abrams, M.; and Scheutz, M. 2024. Automating Dataset Production Using Generative Text and Image Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 1988–1995.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Wei, J.; and Zou, K. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Wu, J.; Antonova, R.; Kan, A.; Lepert, M.; Zeng, A.; Song, S.; Bohg, J.; Rusinkiewicz, S.; and Funkhouser, T. 2023. Tidybot: Personalized robot assistance with large language models. *Autonomous Robots*, 47(8): 1087–1102.