LLM-Augmented Symbolic NLU System for More Reliable Continuous Causal Statement Interpretation

Xin Lian X.LIAN@U.NORTHWESTERN.EDU
Kenneth D. Forbus FORBUS@NORTHWESTERN.EDU
Department of Computer Science, McCormick School of Engineering and Applied Science, Northwestern University, Evanston, IL 60208

Abstract

Despite the broad applicability of large language models (LLMs), their reliance on probabilistic inference makes them vulnerable to errors such as hallucination in generated facts and inconsistent output structure in natural language understanding (NLU) tasks. By contrast, symbolic NLU systems provide interpretable understanding grounded in curated lexicons, semantic resources, and syntactic & semantic interpretation rules. They produce relational representations that can be used for accurate reasoning and planning, as well as incremental debuggable learning. However, symbolic NLU systems tend to be more limited in coverage than LLMs and require scarce knowledge representation and linguistics skills to extend and maintain. This paper explores a hybrid approach that integrates the broad-coverage language processing of LLMs with the symbolic NLU capabilities of producing structured relational representations to hopefully get the best of both approaches. We use LLMs for rephrasing and text simplification, to provide broad coverage, and as a source of information to fill in knowledge gaps more automatically. We use symbolic NLU to produce representations that can be used for reasoning and for incremental learning. We evaluate this approach on the task of extracting and interpreting quantities and causal laws from commonsense science texts, along with symbolic- and LLM-only pipelines. Our results suggest that our hybrid method works significantly better than the symbolic-only pipeline.

1. Introduction

Recent years have seen the dramatic advancement of large language models (LLMs), whose capabilities now span a wide range of tasks, including natural language processing, question answering, and even creative generation (Achiam et al., 2023; Team et al., 2023; Dubey et al., 2024; Abdin et al., 2024; Gunter et al., 2024). Do they genuinely understand natural language, though? LLMs do not engage in formal, structured reasoning in the way humans do: while some studies have drawn connections between LLM mechanisms (e.g., attention and residual streams) and cognitive psychology constructs (Goyal & Bengio, 2022; Yu & Ananiadou, 2023), LLMs fundamentally operate through probabilistic pattern matching rather than principled reasoning procedures (Jiang et al., 2024), which causes inevitable issues like hallucination (Marcus, 2020; Huang et al., 2025). Particularly, it has been shown that LLMs cannot learn certain fundamental semantic properties in lan-

guage understanding, including semantic entailment and consistency as defined in formal semantics (Asher et al., 2023).

Natural language understanding (NLU) systems, predating the widespread adoption of LLMs, integrate multiple aspects of language processing – syntax, semantics, and inference – often through a parser grounded in hand-crafted rules and a formal model of linguistic structure (Winograd, 1972). More advanced symbolic, and in some cases hybrid, systems include NL-Soar (Rubinoff & Lehman, 1994; Lindes & Laird, 2017), Companion NLU (CNLU, Tomai & Forbus, 2009), and the Never-Ending Language Learner (Mitchell et al., 2018). These knowledge-based systems use structured, discrete, and inspectable representations, enabling interpretations that are explainable and supporting reasoning. Despite these advantages, symbolic NLU systems face persistent challenges: natural language is complex and constantly evolving, while the coverage of their lexicons and semantic knowledge bases is inherently limited, sometimes incorrect, incomplete, outdated, or unsuitable for certain domains. This restricts their applicability and necessitates significant manual effort to expand rule sets and knowledge bases according to the system's ontological framework. Although LLMs lack explicit reasoning in the symbolic sense, they still dominate language processing due to their ability to generate text through statistical pattern matching and their training on vast amounts of text. As such, they present a promising complementary resource: filling gaps in symbolic systems' knowledge bases.

In this paper, we present our ongoing work on integrating LLMs into the CNLU system to expand its capabilities for language understanding. Our focus here is on the domain of comparative analysis questions, using texts from the QuaRTz dataset (Tafjord et al., 2019b). We show how LLMs can be used to translate the diverse natural language forms of wild text into simplified language that CNLU can more easily process. We also show how LLMs can be used to extend the formal semantics that CNLU uses, expanding its lexicon. We compare this hybrid approach with pure CNLU and LLM baselines over the same texts, showing that the hybrid version outperforms the pure CNLU baseline and becomes more comparable to the pure LLM baseline. As this is our initial attempt to help CNLU expand semantics and achieve better performance in qualitative reasoning tasks with LLMs, we then analyze some possible ways to improve as future steps.

2. Background

2.1 CNLU

The Companion Natural Language Understanding System (CNLU) (Tomai & Forbus, 2009) is a rule-based semantic parser. It is grounded in the NextKB[1] knowledge base (KB), which uses the OpenCyc ontology, KB contents, and microtheory structure (microtheories provide a means of contextualization). NextKB includes a large open-license English lexicon, an integration of FrameNet (Baker et al., 1998; Fillmore et al., 2001) and Discourse Representation Theory, as well as support for spatial, qualitative, and analogical reasoning and learning.

CNLU parses texts using the Allen Trains chart parser (Allen, 1995), processing tokens left-to-right and bottom-up while incrementally matching grammar rules to generate parse tree candidates. Lexical features are specified via facts in the KB and augmented with semantic translation (sem-

1. For more details, see `https://www.qrg.northwestern.edu/nextkb/index.html`

trans) facts that tie words to the ontology. A semtrans specifies the semantic information a word implies in context. For example, one semtrans for the comparative adjective `cooler` encodes a comparison event whose associated quantity type is `Temperature`, with a word frame type representing the level of force or exertion, and a comparative relation `lessThan`:

```
(fnsemtrans (TheList) Cooler-TheWord Adjective
  (and (isa :EVENT ComparisonEvent) (comparer :EVENT :SUBJECT)
       (compareе :EVENT :NOUN) (comparisonReln :EVENT lessThan)
       (comparisonQtype :EVENT Temperature)) (frame FN_Temperature)
  (bindingTemplate (TheList)) (groupPatterns (TheSet)))
```

These semtrans play a direct role in generating formal interpretations of each lexical token in a given sentence in choice sets, based on their roles computed by the parser and denoted by keywords (here `:EVENT`, `:SUBJECT`, and `:NOUN`). For example, given the sentence "When particles move more slowly, temperature is lower and an object feels cooler", the lexical token `cooler` would encode a piece of formal representation as part of the semantic interpretation as follows, indicating the comparison event implied:

```
(and (isa cool13002 ComparisonEvent)
 (comparer cool13002 object11166) (compareе cool13002 object11166)
 (comparisonReln cool13002 lessThan) (comparisonQtype cool13002 Temperature))
```

Selecting the appropriate semantic interpretation, or using them to generate the desired logical form in a given context, requires some means for disambiguation and further semantic analysis. As per (Tomai & Forbus, 2009) these further semantic analyses can, via abductive reasoning, serve to also disambiguate at the same time. Two families of methods we have explored are: (1) analogical Q/A training with query cases (Crouse et al., 2018; Wilson et al., 2019; Ribeiro et al., 2019; Crouse, 2021; Ribeiro & Forbus, 2021), which map semantic choices to desired logical forms for specific contexts; or (2) narrative functions (McFate et al., 2014), which construct logical forms based on interpretation rules and grammatical–semantic knowledge, sometimes with additional statistical methods (Ribeiro & Forbus, 2021).

2.2 QP Theory and Linguistic Frames

Qualitative Process (QP) Theory (Forbus, 1984) provides a formal theory for qualitative causal mathematics that includes the kinds of phenomena explored in QuaRTz problems. Specifically, *qualitative proportionalities* are partial information about functional relationships between quantities. For example, `(qprop <temperature of gas> <speed of its particles>)` captures part of the intended meaning of the earlier example. `qprop` implies an increasing monotonic functional dependency, so here a lower speed would license the inference that the temperature is lower, all else being equal. These relationships are partial and need to be combined to construct a complete causal account to reason with. This compositionality of qualitative models is an excellent fit for the compositionality of language. Expressing complex ideas in language typically requires carving them up into pieces so they can be reassembled by the recipient. This makes qualitative modeling a natural approach to aspects of natural language semantics.

For constructing qualitative knowledge from language, Kuehne (2004) demonstrated that QP theory constructs can be represented in a frame-based format compatible with the semantic frames of FrameNet, highlighting a natural alignment between the organization of physical knowledge in

QP theory and the structure of frame semantics. Subsequent work extended this frame notion to a broader range of qualitative models and narrative functions for extracting them from language (e.g. McFate et al., 2014; Forbus & Hinrichs, 2020). We illustrate by example, given the statement "The Sun is large", CNLU with narrative functions produces the following quantity frame, where `-DV` means the discourse variable created during the analysis:

```
((QuantityFrame4126
  (isa QuantityFrame4126 QuantityFrame) (quantityEntity QuantityFrame4126 TheSun)
  (quantityEntityDV QuantityFrame4126 TheSun) (quantityType QuantityFrame4126 Volume)
  (quantityValue QuantityFrame4126 (HighAmountFn Volume))))
```

Similarly, to encode a qualitative proportionality between quantities, narrative functions construct an influence frame with slots for a constrainer (cause) quantity, a constrained (effect) quantity, and a sign indicating the direction of change. So in this case, QP frames would be the desired logical forms when trying to generate interpretations that describe the causal qualitative model.

2.3 QuaRTz

One of the earliest natural language datasets focused on qualitative relationship reasoning was *QuaRel* (Tafjord et al., 2019a), which contains 2,771 multiple-choice story questions paired with logical forms. However, these questions are limited to a predefined set of qualitative relations and vary only in narrative context and subject matter. To address this limitation, Tafjord et al. (2019b) later introduced *QuaRTz*, the first open-domain dataset for reasoning about textual qualitative relationships. QuaRTz includes 3,864 crowdsourced, situated questions, each annotated with the properties being compared and the corresponding grounding fact, which is a natural language sentence conveying the qualitative knowledge required to answer the question. Unlike QuaRel, QuaRTz is not restricted to a fixed inventory of qualitative relations. For example, the questions:

> *(1) If Mona lives in a city that begins producing a greater amount of pollutants, what happens to the air quality in that city? (A) increases (B) decreases*
> *(2) As Cindy swam closer to the ocean surface, it got lighter and ___ (A) warmer (B) cooler*

are supplemented with respective grounding facts that can be used in finding an answer:

> *(1) More pollutants mean poorer air quality.*
> *(2) Water density increases as salinity and pressure increase, or as temperature decreases.*

A grounding fact may imply comparisons in two ways: (1) between different quantities treated as entities, whose amounts or degrees are compared (as in Question 1), or (2) between quantities associated with the same entity set (as in Question 2). Moreover, a grounding fact such as that for Question 2 can mention multiple constrainer and constrained quantities, thereby encoding several causal relations, even though not all of them are directly relevant to answering the question.

2.4 Previous Work in Differential Qualitative Analysis

One promising way to solve qualitative reasoning problems like the ones in QuaRel or QuaRTz is via comparative analysis (Weld, 1988), which uses qualitative representations to infer the causal

consequences of differences, either introduced hypothetically within a system or observed between two distinct physical systems. In particular, differential qualitative analysis (DQA) (Weld, 1988) applies a set of inference rules to compute relative values across system descriptions based on specified differences. Klenk et al. (2005) further demonstrated that DQA can be generalized by employing analogical mappings to align the compared systems automatically. An LLM might also be used in solving such problems, but it would not be based on an explicit model of some phenomenon, whereas differential analysis with DQA is always based on an explicit qualitative causal model.

Crouse et al. (2018) was the first to apply DQA, augmented with analogical mappings, to solve problems from the QuaRel dataset, using mappings between structured situation descriptions derived from natural language. Inspired by the original QuaRel design, which models comparison questions as involving two "worlds" (Tafjord et al., 2019a), each represents one change in a particular quantity from one or more entities. Crouse constructed microtheories for each compared "world", embedding their qualitative representations accordingly. However, the "two-world" separation is ultimately a lazy abstraction – the questions are schematically constructed around a fixed set of qualitative relationships, assuming two distinct quantities are always compared, whereas it breaks down with more complex questions which may involve more entities, or quantities compared. Ultimately, effective language understanding should be grounded in the semantic meanings of the entities within a sentence, rather than reduced to an artificial partition into two "worlds".

2.5 Augmenting LLMs with Symbolic Systems

The texts in QuaRel and QuaRTz are free-form, everyday texts, while the syntactic and lexical coverage of CNLU is much narrower. Prior evidence suggests that incorporating LLMs can help. For example, Ribeiro & Forbus (2021) integrated BERT into Companion to assist with disambiguation in NLU for learning by reading from Simple English Wikipedia articles. They fine-tuned BERT on FrameNet data to serve as a frame classifier and further adapted it to assess fact plausibility by training it on examples of correct and incorrect textual statements.

Why not replace CNLU with an LLM trained to produce NextKB representations? First, modern ontologies are huge, which would require massive amounts of training data. For example, NextKB includes over 80,000 concepts and over 20,000 predicates, and its 700,000+ facts are distributed across over 1,300 microtheories. Second, ontologies (like language) evolve incrementally, which is incompatible with the batch learning method of LLMs. Finally, the inspectability and debuggability of symbolic systems are an important benefit. Thus, we have been exploring several alternatives. Nakos & Forbus (2023) proposed using LLMs to simplify text, since text-to-text tasks are natural for LLMs, and regularizing the language input should increase CNLU's coverage. A second use of LLMs treats them as an informant, getting information from them in language that the receiving system must turn into knowledge. For example, Kirk et al. (2022) used prompt engineering to guide an LLM in producing text that could be read by a Soar agent under human oversight, and later reduced this oversight through a framework that combined LLMs with tree search (Kirk et al., 2024). Later, Wray et al. (2024) proposed an LLM-augmented framework, the Cognitive Task Analyst, assigning LLMs more specific, curated roles in generating formal representations for defining problem spaces. In this paper, we adopt a similar philosophy: LLMs serve curated, well-defined

roles in supporting semantic representations generation, thereby improving language understanding of the NLU system.

3. Task Description

In this paper, we focus on understanding the grounding facts for each QuaRTz question, namely the continuous causal statements describing causal relationships between quantities that are relevant to answering the question. More specifically, with LLM assistance, our hybrid approach aims to extract the quantities mentioned in each grounding fact, along with the influence signs (i.e., directions of change) for each quantity. For example, given the grounding fact "When particles move more slowly, temperature is lower and an object feels cooler", we expect the system to produce a formal representation identifying the quantities and their corresponding influence signs, such as `(speed -) (temperature -)` , and such compact representations are acceptable under the assumptions outlined in Section 2.3. Once the system can generate appropriate QP frames, either indicating an influence relation or an ordinal relation between compared entities for a given quantity, we can convert these frames into readable assertions and directly identify the quantity types and influence signs, as shown in the following. In this paper, we primarily focus on generating correct ordinal frames (which imply comparison events).

```
;; Quantity frames for quantity type ‘Speed‘
(QuantityFrame1
  (isa QuantityFrame1 QuantityFrame) (quantityType QuantityFrame1 Speed)
  (quantityEntity QuantityFrame1 particle1))
(QuantityFrame2
  (isa QuantityFrame2 QuantityFrame) (quantityType QuantityFrame2 Speed)
  (quantityEntity QuantityFrame2 (GapFn :NOUN)))
;; Quantity frames for quantity type ‘Temperature‘
(QuantityFrame3
  (isa QuantityFrame3 QuantityFrame) (quantityType QuantityFrame3 Temperature)
  (quantityEntity QuantityFrame3 particle1))
(QuantityFrame4
  (isa QuantityFrame4 QuantityFrame) (quantityType QuantityFrame4 Temperature)
  (quantityEntity QuantityFrame4 (GapFn :NOUN)))
;; Ordinal frames -- interpreted as comparison events
(OrdinalFrame1
  (isa OridinalFrame1 OrdinalFrame) (OrdinalReln OrdinalFrame1 lessThan)
  (quantity1 OrdinalFrame1 QuantityFrame1) (quantity2 OrdinalFrame1 QuantityFrame2))
(OrdinalFrame2
  (isa OridinalFrame2 OrdinalFrame) (OrdinalReln OrdinalFrame2 lessThan)
  (quantity1 OrdinalFrame2 QuantityFrame3) (quantity2 OrdinalFrame2 QuantityFrame4))
```

4. Methodology

The set of quantities that people deal with is vast, larger than CNLU's lexicon. Hence we start by describing how we dynamically expand the lexicon by using information extracted by an LLM on the QuaRTz training set. Then describe how QP frames are extracted for each grounding fact, and the pipelines we used for comparison: pure-LLM, pure-CNLU, and the hybrid approach, where CNLU is assisted with LLM in either rephrasing original grounding facts or expanding the lexicon. Finally, we explain the metrics used in the evaluation. Note that in our lexicon diagnosis, lexicon

expansion, and experiments, Phi-4 (Abdin et al., 2024) was used where an LLM is needed, because it is relatively small and, in our observations, produces more stable outputs[2].

4.1 Lexicon Diagnosis and Expansion

Figure 1 illustrates the procedure for checking and expanding the lexicon. The process is incremental and sequential—it processes one comparative adjective at a time rather than in batches, so each new evaluation consults the KB as updated by any semtrans added in earlier steps. In the following, we describe each step in detail: rephrased text generation, semtrans diagnosis, new semtrans construction, and determining frame-quantity relevance.

4.1.1 Rephrased Texts Generation

Given a grounding fact, we first prompt an LLM to generate rephrased sentence(s) in a fixed comparative structure designed to be easier for CNLU to parse and interpret:

```
[Noun] that is/has [comparative adjective] / [more/less quantity-noun] is/has
[comparative adjective] / [more/less quantity noun] than [Noun].
```

In this way, given the original grounding fact example "If the gas is cooled, the particles will move more slowly, because they will have less energy", the LLM produces:

> *Particles that are in cooler gas move more slowly than particles.*
> *Particles that are in cooler gas have less energy than particles.*

4.1.2 Semtrans Diagnosis

For each rephrased text derived from the original grounding fact, CNLU first tokenizes the sentence and identifies comparative adjectives using the lexicon. For each comparative adjective, CNLU then searches the KB for any existing semtrans entries. If none are found, the system proceeds directly to a new semtrans generation. Otherwise, it filters the retrieved semtrans to retain only those that imply a comparison event (see Section 2.1 for an example). If no semtrans remain after this filtering, the system likewise proceeds to new semtrans generation.

When one or more semtrans introduce a comparison event for the comparative adjective, the system asks the LLM to assess whether each semtrans is sufficiently relevant to the given grounding fact, with particular attention to the quantity type and frame type it specifies. If none of the semtranses are relevant, the system proceeds to construct new semtranses.

4.1.3 New Semtrans Construction

Figure 2 outlines the process for constructing a new semtrans for a comparative adjective deemed insufficiently relevant to the grounding fact. The system first queries the lexicon for the adjective's root form (e.g., *dense* for *denser*) and runs a similar diagnostic process: (1) If the root form has semtrans in the KB, continue; (2) Retain the semtrans of the root form that includes a qualitative

2. Here, "stable" means fewer explanations or side notes; we found that larger LLMs tend to generate more irrelevant content in addition to the desired output, a trend also noted by others (Zhou et al., 2024).

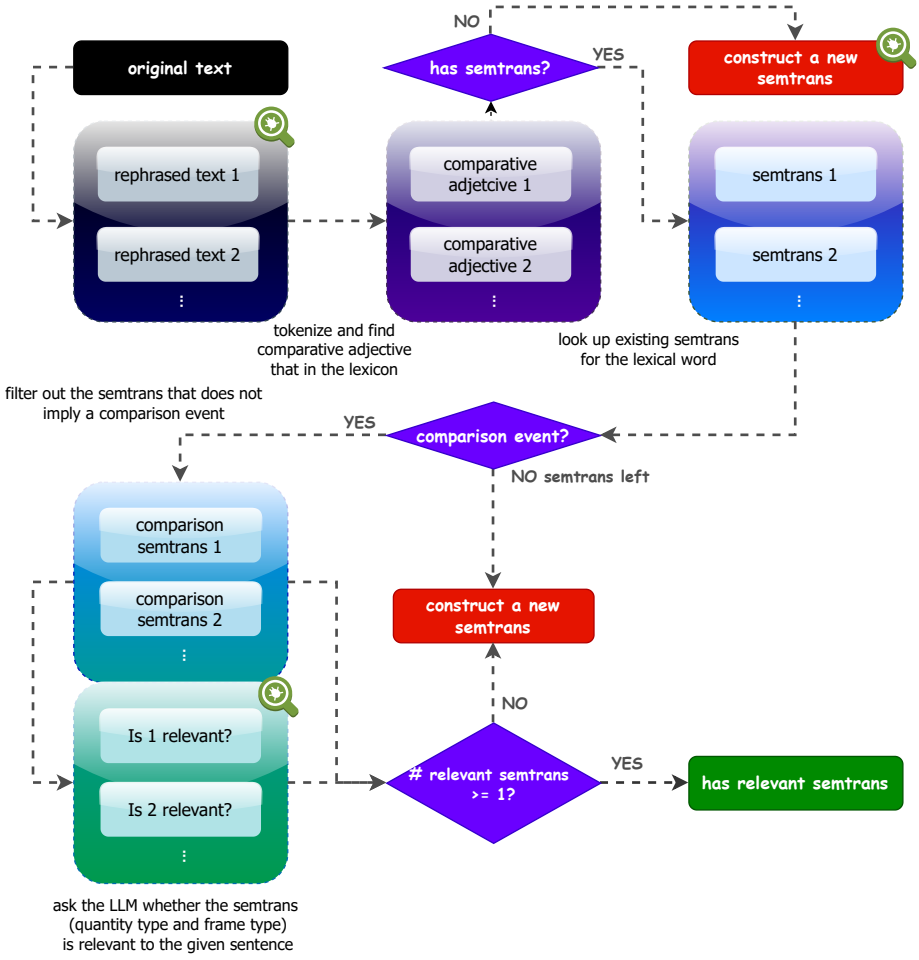

Figure 1. Procedure for diagnosing an existing semtrans for a comparative adjective, together with its associated grounding fact under inspection. The diagnosis results in either (1) identifying at least one relevant semtrans from the existing knowledge base, or (2) constructing a new semtrans if: no semtrans exists for the adjective, no existing semtrans encodes a comparison event, or no existing semtrans is sufficiently relevant to the adjective as used in the given text. The dark green magnifying glass icon in the upper-right corner of a step indicates that an LLM assists the step.

value of a particular type of quantity (e.g. `(HighAmountFn Density)` for *dense*); (3) Look for the quantity type given the qualitative value (e.g. `Density` for *dense*); (4) Ask the LLM whether the quantity type is relevant given the comparative adjective's use in the original grounding fact; (5) Adopt the word frame and quantity type in the semtrans that is relevant. So as an example, given the comparative adjective *denser*, the system retrieves its root form *dense* and uses its relevant semtrans to collect the word frame and quantity types (`FN_Misc` and `Density`).

If it fails to retrieve any relevant semtrans from the root form, the system turns to its antonym (obtained via LLM, as the lexicon does not store antonym information) and applies the same diagnostic process. For instance, to construct a semtrans for *closer*, assuming its root form *close* does not have any relevant semtrans given some grounding fact, the system then inspects its antonym *farther* and adopts the most relevant word frame–quantity type pair (`FN_Gradable_proximity` and `Distance`) after evaluation.

Finally, suppose it still fails to find any relevant semtrans from the antonym form. In that case, it queries the root form of the antonym from the lexicon, then follows the same process as before to see whether a relevant word-quantity-frame pair exists in the KB. For instance, if it cannot find any relevant word-quantity-frame pair for original comparative adjective *closer* after inspecting its root form *close* and antonym *further*, it then inspects the root form of its antonym *far*.

Once the appropriate word frame and quantity types are identified – whether from the root form, the antonym, or the root form of the antonym – the system queries the LLM for the influence sign, given the quantity type and comparative adjective. With the word frame type, quantity type, and influence sign determined, the new semtrans is created and added to the KB, making it available for future semtrans diagnoses.

4.1.4 Determining Frame-Quantity Relevance

Given a semtrans, specifically its word frame and quantity type, the system queries the LLM to determine whether they are relevant to a given grounding fact, taking into account the original use of the comparative adjective in that fact. For example, suppose the grounding fact is "If the current passing through the circuit breaker increases, the electromagnet becomes stronger", where the comparative adjective is *stronger*, in the case when its root form (*strong*), its antonym (e.g., *weaker*), or the root form of its antonym (e.g. *weak*) is inspected. The system has access to candidate frame-quantity-type pairs such as `(Fn_Usefulness Effectiveness)`, `(FN_Expertise Expertise)`, and `(FN_Level_of_force_exertion Strength)`[3]. The LLM is then asked, for each pair, whether it is relevant to the comparative adjective in the grounding act. In this case, the first two pairs should be judged irrelevant, while the last pair is relevant. Note that one alternative way to determine the relevance between a frame-quantity-type pair and the grounding fact is to compute a "relevance score" quantitatively. This could be done by combining semantic similarity, comparative cues, and part-of-speech-based comparative patterns, and then evaluating whether an auto-generated sentence based on the frame-quantity-type pair lies sufficiently close to the grounding fact in the sentence embedding space (Reimers & Gurevych, 2019; Ethayarajh, 2019). However, this approach is less reliable: semantic relevance may not be fully captured by proximity in the embedding space alone, and our preliminary experiments showed that its performance was markedly lower than that of determining relevance by directly querying the LLM.

3. In CNLU, these frame or quantity types, which are raw entities in the KB, are converted to natural language by a built-in natural language generation system to make sure the LLM understands what the entities mean. For example, given frame type `FN_Level_of_force_exertion`, the query to the LLM would just mention it as "level of force exertion".

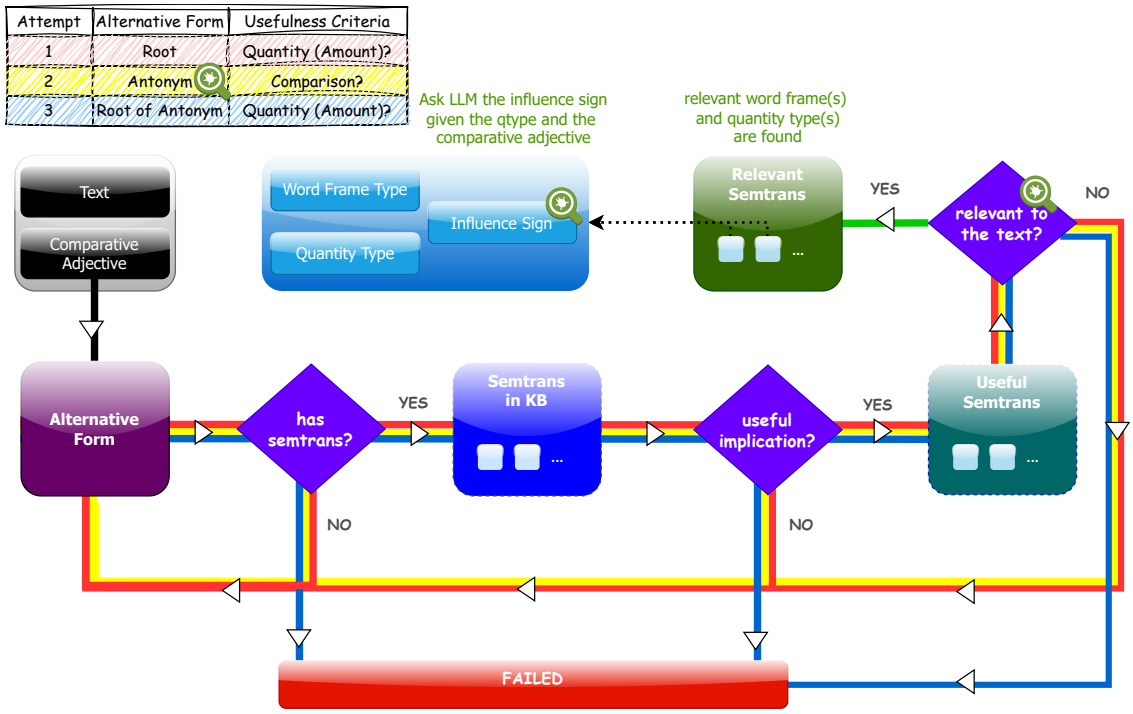

Figure 2. Procedure for constructing a new semtrans for a comparative adjective deemed insufficiently relevant to the given grounding fact. The system begins by searching the relevant semtranses of the adjective's root form (red path), followed by the relevant semtranses of its antonyms (yellow path), and finally the relevant semtranses of each antonym's root form (blue path). If any search yields semtranses relevant to the grounding fact, the procedure stops and returns the identified quantity type(s) and frame type(s), along with the corresponding influence sign(s) obtained via the LLM. Using these results, the system generates new semtrans(es). If no relevant semtranses are found, the procedure terminates without generating new ones. A dark green magnifying glass icon in the upper-right of any step indicates LLM assistance.

4.2 Pipelines

In our experiments, we evaluated the following pipelines:

- **Pure LLM**: We used Phi-4 to generate the desired logical forms directly from curated prompts containing a few illustrative examples.
- **Pure CNLU**: We used CNLU only – it does not use either the updated lexicon or the LLM-based sentence rephrasing. It generates semantic interpretations directly from the original text, produces QP frames based on these interpretations with narrative functions (Forbus et al., 2020), and then the QP frames are converted into readable assertions from which the quantity types and their influence signs can be extracted.

- **Hybrid**: The hybrid pipelines generate the desired logical forms by CNLU with the assistance from Phi-4 – so CNLU with either the updated lexicon, the sentence rephrasing with Phi-4, or both.

The pure CNLU and hybrid pipelines are expected to produce logical forms such as:

```
(‘‘When particles move more slowly, temperature is lower and an object feels cooler
    .’’ (Speed -) (Temperature -))
```

Here, each output includes the original grounding fact along with quantity–sign pairs. In contrast, the pure LLM pipeline is expected to produce logical forms in the following format:

```
(‘‘When particles move more slowly, temperature is lower and an object feels cooler
    .’’ (cause speed) (effect temperature) (cause-sign -) (effect-sign -))
```

Note that, when there is more than one cause or effect quantity type, the number of signs should match the number of quantity types in the logical form, like the example in the following:

```
(‘‘Bigger stars produce more energy, so their surfaces are hotter.’’
 ((cause size) (effect amount-energy temperature) (cause-sign +) (effect-sign + +))
```

4.3 Evaluation Metrics

For a tested grounding fact i, let the gold quantity-sign specification be Q_i and $s_i : Q_i \to \mathcal{S}$, where Q_i is the set of gold quantity types and $\mathcal{S} = \{+, -, None\}$[4] the set of influence signs. Each evaluated pipeline outputs the set of quantity types $\hat{Q}_i$ with their respective influence signs $\hat{s}_i : \hat{Q}_i \to \mathcal{S}$.

To allow partial credit for near quantity type matches, while never rewarding signs when the quantity type is wrong, we define for each gold $q \in Q_i$, a matched prediction $\hat{q} \in \hat{Q}_i \cup \varnothing$, and a quantity match weight (in the experiments we set $\alpha = 0.5$):

$$w_i(q) = \begin{cases} 1, & \text{exact, or ontology-equivalent match} \\ \alpha \in (0,1), & \text{partially relevant} \\ 0, & \text{not match} \end{cases}$$

Then the pipeline will have the following metrics regarding the tested grounding fact:

- **Quantity Coverage (QC) Score**: measures how well the system recovers the intended quantity types (full or partial matches), given by

$$\frac{1}{|Q_i|} \sum_{q \in Q_i} w_i(q)$$

- **Conditional Sign Accuracy (CSA) Score**: defined only when $\sum_{q \in Q_i} w_i(q) > 0$, otherwise reported as N/A; so if a quantity type is not matched, its sign still contributes no credit. It is given by

$$\frac{\sum_{q \in Q_i} w_i(q) \cdot \mathbf{1}[s_i(q) = \hat{s}_i(\hat{q})]}{\sum_{q \in Q_i} w_i(q)}$$

4. If None, the system fails to find the sign for the found quantity type.

Table 1. Metric scores for each pipeline.

Pipeline		QC	CSA	OP
CNLU		16.05	90.00	11.78
Hybrid	Updated Lexicon	22.33	91.05	17.44
	Text Rephrase	26.60	94.44	22.62
	Both	34.67	90.76	30.90
Phi-4		89.86	95.21	87.88

- **Overall Pair (OP) Score**: yields a single scalar per fact that (1) rewards finding the appropriate quantity types (with partial credit possible), and (2) rewards getting the sign correct only when the quantity is (at least partially) correct. It is given by

$$\frac{\sum_{q \in Q_i} w_i(q) + \sum_{q \in Q_i} w_i(q) \cdot \mathbf{1}[s_i(q) = \hat{s}_i(\hat{q})]}{2|Q_i|}$$

For example, consider the grounding fact "When a gas is squeezed into a smaller volume, the particles have less space to move" whose gold quantity–sign pairs are `(Volume -) (SpatialQuantity -)`. Suppose the system predicts `(Volume -) (Size -)`. By assigning a partial quantity weight of 0.5 to `Size` as a near match to `SpatialQuantity` and a quantity weight of 1.0 to `Volume` as an exact match, the QC score becomes

$$\frac{1.0 + 0.5}{2} = 0.75$$

Since the returned signs for both quantity types are accurate, the CSA score is given by

$$\frac{1.0 \cdot 1 + 0.5 \cdot 1}{1.0 + 0.5} = 1.0$$

Finally, the OP score is given by

$$\frac{(1.0 + 0.5) + (1.0 \cdot 1 + 0.5 \cdot 1)}{2 \cdot 2} = 0.75$$

indicating an overall "matching" score. Note that, these metrics assess coverage of the gold quantity-sign pairs rather than the exact equality of the gold and generated quantity-sign pairs. After computing scores for all tested grounding facts, we report averages for each evaluated pipeline, each times 100% as the actual score out of 100; for conditional sign accuracy, we average only over facts with at least one matched quantity (i.e., nonzero denominator).

5. Results

By using 282 distinct grounding facts in the QuaRTz training set, the system constructed 28 new semtrans facts, all of which are reasonable. After that, all pipelines generated the expected logical form, indicating the associated quantity types and respective influence signs, given the 81 distinct grounding facts in the QuaRTz test set. Their metric scores are shown in Table 1.

5.1 LLM-Only

The Phi-4–only pipeline performs strongly on extracting quantity types and their influence signs. In most test cases, its generated logical forms cover the required quantities implied by the grounding facts and assign the correct influence signs to those quantities, yielding high QC and CSA scores. It achieves an OP score of 87.88 (0–100 scale), outperforming the other pipelines on our evaluation.

However, it still fails in some cases that humans find straightforward, and CNLU or hybrid pipelines can solve better. For example, given the test case "Convection on early Earth was faster and so plate tectonics was faster," the relevant quantity types should be the Earth's time or age (cause) and the event rate of plate-tectonic motion (effect). Phi-4 instead returned `convection` and `speed-plate-tectonics`. The former receives no credit because "convection" is a process or entity, and even if it can be wrapped as a quantity type, it would still be irrelevant to the time or age of the Earth; the causal quantity must be the *age of the Earth*. The latter is also not perfect: "speed" typically denotes motion over distance per unit time, whereas the intended construct is an *event rate* (e.g., `EventRate`); this is captured more appropriately by the CNLU or hybrid pipelines. More broadly, this kind of error is consistent with an association-based inference strategy: an LLM tends to surface distributionally "relevant" lexical items (e.g. *convection*, *speed*) that co-occur with cues like *faster*, rather than projecting the text into the KB's quantity ontology and enforcing causal role constraints. In effect, it retrieves salient entities instead of the latent quantities that license the inference (here `Age` or `AGE(Earth)` as cause, and `EventRate` as effect), which has been illustrated as the shortcut or heuristic behavior in language models (Bender & Koller, 2020; McCoy et al., 2019). This also aligns with our metric pattern – high CSA conditional on a matched quantity, but lower QC when entity-quantity disambiguation is required, suggesting probabilistic association rather than quantity-level reasoning in such cases.

The LLM can also introduce occasional structural errors in the logical forms, such as missing signs or fusing multiple quantities into a single term. For instance, the following output omits the effect sign for `depth`:

```
("Older layers are deeper in the Earth, younger layers are closer to the surface."
  ((cause age) (effect depth) (cause-sign +)))
```

In another case, multiple distinct quantity types are merged into an integrated label, leading to inconsistent formal representations:

```
("A gentle slope favours slower flow of surface water, reduces erosion, and
 increases availability of water to plants."
  ((cause slope) (effect speed-erosion-availability-water)
   (cause-sign -) (effect-sign - - +)))
```

Beyond such formatting errors, two practical concerns remain:

- **KB interoperability.** Terms in the generated logical forms often lack a direct mapping to entities in the KB, limiting immediate consumption by symbolic components such as CNLU.
- **Potential data contamination.** QuaRTz was designed to inspect neural models' reasoning abilities, and has been available on the open web since its publication. Hence it is very likely that the entire dataset has been used as part of the training corpus of LLMs. System performance may therefore depend on prior exposure rather than general language or reasoning capabilities.

Despite these issues, within our setup, the LLM-only pipeline is generally much more reliable at identifying relevant quantity types than the CNLU-only and hybrid variants.

5.2 CNLU-Only and Hybrid

On longer and syntactically complex grounding facts, the CNLU-only pipeline frequently fails to produce full parses or reliable semantic choices, and—even when parses are available – often misses quantity types implied by the sentence. Nevertheless, it attains a QC score of 16.05 and, despite relatively strong sign handling on the quantities it does find (high CSA score), its overall score remains modest with the OP score 11.78, reflecting low coverage.

As mentioned, augmenting CNLU with Phi-4 improves performance via two mechanisms: **(i) Lexicon update.** With Phi-4 assisting lexicon expansion (while retaining CNLU's original syntactic parsing), the pipeline reaches a QC score of 22.33 and an OP score of 17.44. **(ii) Text rephrasing.** When Phi-4 rephrases grounding facts into more CNLU-amenable forms, the pipeline attains the QC score 26.60 and a high CSA score 94.44, yielding the OP score 22.62. **(iii) Both lexicon update and rephrasing.** Combining both interventions yields the best hybrid results: A QC score of 34.67 and an OP score of 30.90. In other words, the hybrid pipeline correctly covers more than one-third of the required quantity types, while remaining imperfect on some signs.

Still, there are some limitations of the hybrid approach that hinder it from having comparable performance with Phi-4:

- **Incomplete coverage of comparative determiners.** The current lexicon diagnosis and expansion targets comparative adjectives that directly signal comparisons over their implied quantity types. Degree determiners and noun-modifying comparatives (e.g., *more*, *less*, *higher ⟨quantity⟩*, *lower ⟨quantity⟩*) are not yet handled well in context, and such cases are common in QuaRTz and central to describing quantity change.

- **Residual complexity after rephrasing.** Some LLM rephrasings remain multi-clausal or otherwise difficult for CNLU to parse or to map to QP frames. For example, rephrasing "Friction causes the molecules on rubbing surfaces to move faster, so they have more energy. This gives them a higher temperature, and they feel warmer." as "Surfaces that are rubbed together have molecules with more energy than surfaces that are not rubbed together." still leaves structures that challenge CNLU's narrative functions.

- **Occasional irrelevant rephrasings.** Despite our having tried to avoid generations of such texts from Phi-4, it sometimes still produces meta-level or explanatory text that is off-task, or plausible-sounding rephrasings that are semantically unrelated to the original fact, partially due to its hallucinations.

- **Over-rephrasing of simple inputs.** Not all grounding facts should be rephrased. For instance, "More people need more resources" is already canonical for CNLU; rephrasing it to "People that are more in number need more resources than people" distorts the original meaning and harms downstream analysis.

Taken together, these results show that LLM assistance can substantially improve CNLU's ability to identify quantities and their influence signs from raw grounding facts. With better coverage of comparative determiners, stricter controls on rephrasing, and continued lexicon curation, the hybrid approach has clear headroom to approach LLM-only performance while offering greater interpretability and controllability, and without requiring extensive task-specific training data.

6. Conclusions and Future Work

In this paper, we presented our efforts to integrate an LLM with the symbolic NLU system CNLU for generating desired logical forms and interpreting continuous quantity changes from continuous causal statements. In our approach, LLMs are used to rephrase original texts into a regularized format that is easier for CNLU to parse, and to play targeted roles in updating the semantic knowledge base. These roles include supplying lexical information (e.g., antonyms) and semantic information (e.g., the influence sign of a quantity implied by a comparative adjective), as well as supporting statistical methods that use embedding spaces and external lexicons to determine whether a quantity is semantically related to the given text. Our results show that this hybrid method significantly outperforms a purely symbolic NLU system, providing broader language coverage while preserving the advantages of inspectability, support for high-precision reasoning, and incremental rapid learning that symbolic systems provide.

This paper is an initial exploration of incorporating LLMs into the interpretation of natural language texts within a specific reasoning domain (qualitative reasoning) using a massive knowledge base with complex, hand-curated ontologies. Potential future work spans several directions. First, to better handle comparative adjectives that modify the degree or amount of a noun or quantity (e.g., "more water", "less acceleration"), we need encoding strategies beyond semtrans supplementation, for example, more comprehensive grammar rules that directly capture such expressions. Second, the semtrans update procedure should be expanded to incorporate richer methods for identifying potential quantity types. Third, rephrased sentence generation needs to be made more robust. One possible approach is to have LLMs generate structured, fine-grained sentence elements, which CNLU could then assemble into complete sentences according to defined rules. Finally, once the hybrid NLU approach achieves more reliable performance in generating relevant interpretations of quantities, the targeted quantities diagnosed during the semtrans process could be used to specify the quantity types that the NLU system should map to. Possible ways include analogical Q/A training, using query cases constructed from rephrased texts paired with their logical forms, including target quantities only, or using a variation-set like method to invoke an LLM to generate more semantically similar but syntatically different sentences mapping to the logical forms we want.

Acknowledgements

This research was supported by the Office of Naval Research. Additionally, we extend our sincere appreciation to individuals at Qualitative Reasoning Group who generously shared their invaluable insights and feedback on this work, notably Constantine Nakos, Wangcheng Xu, Zoie Zhao, Omar Khater, and Jiahong Zheng.

References

Abdin, M., et al. (2024). Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Achiam, J., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Allen, J. (1995). *Natural language understanding*. Benjamin-Cummings Publishing Co., Inc.

Asher, N., Bhar, S., Chaturvedi, A., Hunter, J., & Paul, S. (2023). Limits for learning with language models. *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (* SEM 2023)* (pp. 236–248).

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The berkeley framenet project. *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Bender, E. M., & Koller, A. (2020). Climbing towards nlu: On meaning, form, and understanding in the age of data. *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5185–5198).

Crouse, M. (2021). *Question-answering with structural analogy*. Doctoral dissertation, Northwestern University.

Crouse, M., McFate, C., & Forbus, K. (2018). Learning from unannotated qa pairs to analogically disambiguate and answer questions. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Dubey, A., et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ethayarajh, K. (2019). How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 55–65).

Fillmore, C. J., Wooters, C., & Baker, C. F. (2001). Building a large lexical databank which provides deep semantics. *Language, Information and Computation: Proceedings of The 15th Pacific Asia Conference: 1-3 February 2001, Hong Kong* (pp. 3–26). Waseda University.

Forbus, K. D. (1984). Qualitative process theory. *Artificial intelligence*, *24*, 85–168.

Forbus, K. D., Hinrichs, E., Crouse, E., & Blass, J. (2020). Analogies versus rules in cognitive architecture. *Proceedings of Advances in Cognitive Systems*.

Forbus, K. D., & Hinrichs, T. (2020). Unifying instance-level and type-level qp frames for natural language understanding. *Proceedings of the 33rd International Workshop on Qualitative Reasoning*.

Goyal, A., & Bengio, Y. (2022). Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, *478*, 20210068.

Gunter, T., et al. (2024). Apple intelligence foundation language models. *arXiv preprint arXiv:2407.21075*.

Huang, L., et al. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*, 1–55.

Jiang, B., Xie, Y., Hao, Z., Wang, X., Mallick, T., Su, W. J., Taylor, C. J., & Roth, D. (2024). A peek into token bias: Large language models are not yet genuine reasoners. *arXiv preprint arXiv:2406.11050*.

Kirk, J. R., Wray, R. E., Lindes, P., & Laird, J. E. (2022). Improving language model prompting in support of semi-autonomous task learning. *arXiv preprint arXiv:2209.07636*.

Kirk, J. R., Wray, R. E., Lindes, P., & Laird, J. E. (2024). Improving knowledge extraction from llms for task learning through agent analysis. *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 18390–18398).

Klenk, M., Forbus, K. D., Tomai, E., Kim, H., & Kyckelhahn, B. (2005). Solving everyday physical reasoning problems by analogy using sketches. *PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE* (p. 209). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Kuehne, S. E. (2004). *Understanding natural language descriptions of physical phenomena*. Northwestern University.

Lindes, P., & Laird, J. E. (2017). Cognitive modeling approaches to language comprehension using construction grammar. *AAAI Spring Symposia*.

Marcus, G. (2020). The next decade in ai: four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*.

McCoy, R. T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3428–3448).

McFate, C., Forbus, K., & Hinrichs, T. (2014). Using narrative function to extract qualitative information from natural language texts. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Mitchell, T., et al. (2018). Never-ending learning. *Communications of the ACM*, *61*, 103–115.

Nakos, C., & Forbus, K. D. (2023). Using large language models in the companion cognitive architecture: A case study and future prospects. *Proceedings of the AAAI Symposium Series* (pp. 356–359).

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982–3992).

Ribeiro, D., Hinrichs, T., Crouse, M., Forbus, K., Chang, M., & Witbrock, M. (2019). Predicting state changes in procedural text using analogical question answering. *7th Annual Conference on Advances in Cognitive Systems*.

Ribeiro, D. N., & Forbus, K. (2021). Combining analogy with language models for knowledge extraction. *3rd Conference on automated knowledge base construction*.

Rubinoff, R., & Lehman, J. F. (1994). Real-time natural language generation in nl-soar. *Proceedings of the Seventh International Workshop on Natural Language Generation*.

Tafjord, O., Clark, P., Gardner, M., Yih, W.-t., & Sabharwal, A. (2019a). Quarel: A dataset and models for answering questions about qualitative relationships. *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 7063–7071).

Tafjord, O., Gardner, M., Lin, K., & Clark, P. (2019b). Quartz: An open-domain dataset of qualitative relationship questions. *arXiv preprint arXiv:1909.03553*.

Team, G., et al. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Tomai, E., & Forbus, K. D. (2009). Ea nlu: Practical language understanding for cognitive modeling. *FLAIRS*.

Weld, D. S. (1988). Theories of comparative analysis.

Wilson, J. R., Chen, K., Crouse, M., Nakos, C., Ribeiro, D. N., Rabkina, I., & Forbus, K. D. (2019). Analogical question answering in a multimodal information kiosk. *Proceedings of the Seventh Annual Conference on Advances in Cognitive Systems*.

Winograd, T. (1972). Understanding natural language. *Cognitive psychology*, *3*, 1–191.

Wray, R. E., Kirk, J. R., & Laird, J. E. (2024). Eliciting problem specifications via large language models. *arXiv preprint arXiv:2405.12147*.

Yu, Z., & Ananiadou, S. (2023). Neuron-level knowledge attribution in large language models. *Conference on Empirical Methods in Natural Language Processing*. From `https://api.semanticscholar.org/CorpusID:266362692`.

Zhou, L., Schellaert, W., Martínez-Plumed, F., Moros-Daval, Y., Ferri, C., & Hernández-Orallo, J. (2024). Larger and more instructable language models become less reliable. *Nature*, *634*, 61–68.