Exploring Hybrid Model Formulation Strategies for Qualitative Reasoning

Kenneth D. Forbus
Qualitative Reasoning Group, Northwestern University
forbus@northwestern.edu

Abstract

One of the core challenges of qualitative reasoning is *model formulation*, building a formal model amenable to qualitative reasoning from situations expressed in everyday modalities (e.g. language, sketching, vision). This paper explores using large language models (LLMs) to help with this process, examining different tradeoffs. We use AI2's QuaRel dataset of comparative analysis problems described in natural language as a testbed. We start with identifying some fundamental tradeoffs, followed by a functional decomposition of QuaRel model formulation to provide context, examining four operations in detail. This paper represents work in progress. We end by summarizing our conclusions so far and outlining plans for future work.

1 Introduction

One of the barriers to the widespread adoption and application of qualitative reasoning is that most qualitative models end up being constructed by hand. This can involve directly working with predicate calculus representations, or with visual languages that simplify the practicalities of dealing with formal languages (e.g. GARP (Bredeweg et al. 2009), Betty's Brain (Leelawong, K. & Biswas, 2008), VModel (Forbus et al. 2004)). The visual languages still have a learning curve, simply not as steep as writing predicate calculus, at the cost of being more limited in the kinds of models they can express. The original situation must be apprehended by the (human) model-builder using natural modalities (e.g. language, sketching, vision), at which humans excel. Progress in other areas of AI and cognitive science are providing new capabilities for building systems that are capable of processing natural modalities, even if not in entirely human-like ways. This paper explores how these new capabilities might be incorporated into qualitative reasoning systems.

Our goal is to automate *model formulation* (Forbus, 2019), and the symmetric process of *model interpretation* (e.g. mapping the results of qualitative reasoning into natural modalities). Prior work on model formulation has mostly focused on determining what additional assumptions are needed, even when starting with formal inputs (e.g. perspective, granularity, subsystems, etc.). Very few efforts have used natural language for expressing scenario models (e.g. Crouse et al. 2018b). One of the most ambitious is Suzuki & Yoshioka's (2024) demonstration of the use of an LLM to construct a qualitative model of a harmonic motion example expressed in natural language, including the automatic extraction of model fragments from the LLM, which is an inspiration for the work described here. Unlike their work, we use a hand-curated domain theory integrated with a language-grounded ontology (Forbus, 2023), to provide stability across operations. Instead we focus on using LLMs to translate problems into more amenable forms.

We begin by explaining our approach to hybrid natural language understanding, including the LLMs we are using and the Companion Natural Language Understanding system (CNLU) that provides our symbolic natural language understanding capability. The QuaRel dataset used here as a testbed is also summarized. Then we do a functional decomposition of the operations needed to solve comparative analysis problems, focusing on model formulation and interpretation, since the qualitative reasoning itself is straightforward. Each step is analyzed with example-based pilot testing to illustrate the approach and issues. We conclude by discussing plans for future work.

2 Hybrid Natural Language Understanding

Advances in large language models have had a revolutionary impact on the field. LLMs have very broad language coverage, capable of taking in any text and doing something with it. Moreover, reinforcement learning has been used to enable these systems to be instructed in natural language (i.e. prompting). For some applications, LLMs have been used to provide more flexible language processing than has been feasible in the past. This makes them a useful off-theshelf technology. However, there are tradeoffs. The problems of confabulation[1] are well known. But these models are also opaque, and the "chain of thought" techniques do not actually reflect an audit trail of the system's reasoning, making their utility questionable (Shojaee et al. 2025). Worse, when they make mistakes, LLMs are not debuggable, in part because all their learning is batch not incremental.

On the other hand, symbolic NLU systems suffer from gaps in coverage, in terms of lexical, grammatical, and semantic knowledge. However, unlike LLMs, they can be debugged and extended incrementally, often via learning (e.g. Crouse et al 2018a,b; Ribeiro et al. 2019). This has led us to explore hybrids, e.g. for learning by reading (Ribeiro & Forbus, 2021), where one can get the best of both worlds.

Model formulation and interpretation is a particularly promising area for hybridization because of its breadth of phenomena. The massive exposure of LLMs to text during training has provided them with contextual correlations that can be used as a form of statistical reasoning. To be sure, much of commonsense is articulable. But by no means all: There are good reasons why it is often called "tacit knowledge" in the QR community! Our question here is, can we use the capabilities of LLMs to provide breadth, with symbolic NLU to provide a bridge to formal QR, to do better at model formulation and interpretation? This paper represents work in progress on exploring that issue.

2.1 CNLU

The Companion Natural Language Understanding system is a knowledge-rich general purpose symbolic language system. It uses a highly modified version of Allen's (1994) TRAINS parser, e.g. it can do reasoning calls to the knowledge base during parsing. Semantics is handled by mappings to the OpenCyc ontology, mediated by FrameNet[2]. Discourse Representation Theory (Kamp & Reyle, 1993) is used to handle complex conditionals, quantification, quotation, and counterfactuals. It is integrated into the Companion cognitive architecture (Forbus & Hinrichs, 2017), and is used in two deployed systems (a Kiosk and a SocialBot), as well as for basic research on multimodal reasoning and learning.

2.2 LLMs used

Following (La Malfa et al, 2025), we eschew the commercial LLMs provided as services in order to avoid problems with replicability. The models we use regularly are Phi4, LlaMA 3.3, and Olmo2. We have tried using "reasoning" models, but so far we find that they blather and do not provide better results.

2.3 QuaRel dataset

The QuaRel dataset (Tafjord et al. 2018) provides a large (2,771) set of comparative analysis questions, all binary choices expressed in English. Qualitative proportionalities are provided as well, and systems are expected to already know this information. The complexity is in the broad language used to describe the scenarios. Here are two examples from the QuaRel training set that we will use as running examples to illustrate ideas:

Example 1: Mike was snowboarding on the snow and hit a piece of ice. He went much faster on the ice because _____ is smoother. (A) snow (B) ice"

Example 2: John and Rita are going for a run. Rita gets tired and takes a break on the park bench. After twenty minutes in the park, who has run farther? (A) John (B) Rita

3 Functional decomposition

In general, qualitative reasoning in natural settings involves understanding a problem well enough to formulate an appropriate model, reason with that model, and interpret the results of the reasoning process in terms of the original problem. Problems can be taken as a scenario about which one or more questions are asked or tasks are requested. Comparative analysis questions involve identifying two situations (or aspects of one situation) to be compared. Following (Klenk et al. 2005), we use the Structure-Mapping Engine (SME; Forbus et al 2017) to construct a mapping between the two situations. Mappings indicate what corresponds with what, e.g. snow/ice in the two snowboarding events in Example 1. The causal laws constraining each situation are part of the situation itself. Differential qualitative (DQ) analysis (Forbus, 1984) is a form of comparative analysis where differences between two situations are propagated through causal relationships to infer what other differences follow. Such differences provide the basis for answering questions, here picking the correct multiple-choice option.

Next we examine these steps more closely, illustrating with examples some of the tradeoffs involved. We suspect that the ambiguous nature of natural language will lead to the need for tighter interaction between QR and NLU, and we note such likely backtracking points below.

4 Extracting situations from scenarios

The first step is extracting the situations to be compared from the scenario. By leaving the output of this step in English, it should be a more natural fit with LLM capabilities and provide an intermediate representation that can easily be understood by CNLU. Figure 1 illustrates a prompt that on Example 1, using Phi4, yields

EventA: Mike snowboarding on snow
EventB: Mike snowboarding on ice

[1] Often called, inaccurately, "hallucinations", which a false sensation, versus generating language without regard to factuality, which is confabulation.

[2] https://framenet.icsi.berkeley.edu/

This is exactly what we want, two easily understandable and easily alignable descriptions of situations. Getting to this point can take some experimentation. For instance, using this minor variation of the prompt

I need you to do some intermediate work in breaking down a problem, to be solved by another system. Here's the problem to be analyzed:
<insert problem here>
DO NOT SOLVE THIS PROBLEM. All I want is something much simpler. First, split this scenario into two analogous events that can be compared. Do not make assumptions that would answer the final question. This part of your response should be of the format
EventA: <descriptionA>
EventB: <descriptionB>
You must keep the two event descriptions as simple as possible.
[…rest of the prompt in Figure 4]

Figure 1: Prompt for situation extraction

"First, split this scenario into two analogous events, <DescriptionA> and <DescriptionB>, keeping them as simple as possible. Output these two events by two lines that look like this"

yields instead

EventA: Mike snowboarding on snow
EventB: Mike hitting a piece of ice

This more literal splitting still allows for the possibility of the hitting event being, for example, a collision with a wall of ice instead of a change in snowboarding surface. Why these two different prompts yield such different results is a mystery. Even with the temperature parameter set to zero, LLM responses to the same input can still vary.

Note the admonishments to not solve the problem and not to generate extra information. Alas today's LLMs are trained to be verbose and obsequious chatbots rather than carefully engineered reliable components. It would be an advance for the field to make better instructible components that do not promote anthropomorphizing.

In using an LLM to extract situations like this, we are assuming that whatever model it has of language is sufficient to do human-like rephrasing, preserving the most likely meaning. Unfortunately, the most likely meaning may or may not be represented in what is the most likely language completion. Consider for example this LLM output for Example 2, with the same prompt:

EventA: John continues running for twenty minutes in the park.
EventB: Rita runs for some time, then takes a break on the park bench for the remainder of the twenty minutes.

There is an ambiguity in the problem as stated: Are John and Rita in the park the entire time? Was John already running and Rita joined him for a while? This last interpretation would support the use of "continues" for John's run, but rules out the two starting together unnecessarily – simply saying "John runs for twenty minutes" would preserve our lack of

```
;;; Snowboarding inherits from Movement-TranslationEvent
(def-encapsulated-history TranslationEvent
    :participants ((theObject :type Physob
                              :constraints
                              (objectMoving ?self theObject)))
  :conditions ((isa ?self Movement-TranslationEvent))
  :consequences
  ((qprop ((QPQuantityFn Distance) ?self)
          ((QPQuantityFn Speed) ?self))
   (qprop- ((QPQuantityFn Time-Quantity) ?self)
          ((QPQuantityFn Speed) ?self))))

;;; Surface contact between board and surfaces make Motion-SolidAgainstSolid relevant
(def-encapsulated-history TranslationOnSurfaceEvent
    :participants ((theObject :type Physob
                              :constraints
                              (objectMoving ?self theObject))
                   (theSurface :type PartiallyTangible
                               :constraints
                               (surfaceObject ?self theSurface)))
  :conditions ((isa ?self Motion-SolidAgainstSolid))
  :consequences
  ((qprop- ((QPQuantityFn Speed) theObject)
           ((QPQuantityFn Friction) theSurface))
   (qprop- ((QPQuantityFn Distance) theObject)
           ((QPQuantityFn Friction) theSurface))
   (qprop ((QPQuantityFn Temperature) theObject)
          ((QPQuantityFn Friction) theSurface))
   (qprop- ((QPQuantityFn Friction) theSurface)
           ((QPQuantityFn SurfaceSmoothness) theSurface))))
```

Figure 3: Encapsulated histories implied by Snowboarding

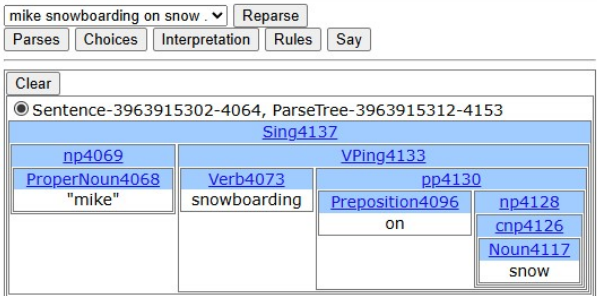
knowledge about how they started. Similarly, the description of Rita's actions, in addition to being longer than convenient for downstream systems, presumes that once Rita stops she continues to rest, as opposed to resting for just a few minutes and then continuing to run. Other versions of the prompt can lead to sentences saying that Rita rests for twenty minutes, which is contradicted by the problem statement. Olmo2 is the most concise but also inaccurate:

EventA: John runs
EventB: Rita rests

We note that these mistranslations would still yield the conclusion that John ran farther. Given the crude signal that multiple choice tests provide, such errors may often not be caught. While more expensive, evaluating the internal model that a system constructs would lead to more confidence in its correctness. If relational internal models are constructed by systems, they could be compared via SME for subsequent evaluation. Let us turn now to model-building.

5 Building QP models for situations

With enough training data and computation, in theory LLMs might be used to automatically translate to logic. But in practice this is daunting. The first problem is that any reasonable commonsense knowledge base provides a large ontology.



● Sentence-3963915302-4064, ParseTree-3963915312-4153

Figure 2: A CNLU parse

For example, the NextKB knowledge base that we use has over 83,000 concepts, 23,000 relations, and over 5,000 logical functions, whose over 700,000 facts are apportioned over 1,400 microtheories. Such coverage would require massive amounts of training data and compute. Hence here we look to CNLU, since we can get an LLM to provide the kind of simple syntax that is easily processed. This also ensures that the language to logic mapping is inspectable, and the conceptual knowledge used in reasoning can be debugged and incrementally extended without massive retraining.

To illustrate how CNLU can be used to set up qualitative models, consider the sentence "Mike snowboarding on snow." Figure 2 illustrates the chart of the syntactic parse. Each word can have multiple interpretations (e.g. here "snow" could refer to the process of snowing or some amount of snow) that are generated as part of the process. How these ambiguities get resolved depends on the task. In some cases domain constraints are used to express preferences, e.g. in conversing about a game, concepts in the game are preferred to other interpretations. If particular kinds of facts are sought due to a task – here, the occurrence of events with continuous aspects – interpretations that include those can be selected by an abduction process built into the reasoner (Tomai & Forbus, 2009).

Figure 3 shows the encapsulated history definitions relevant to snowboarding, inferred via inheritance. That is, the concept of Snowboarding, via a set of superordinate relations, inherits from Movement-TranslationEvent, and the movement of the snowboard against the snow/ice implies a Motion-SolidAgainstSolid event. Recall that encapsulated histories are schemas that represent the occurrence of processes (Forbus, 1984). We use encapsulated histories to capture the qualitative semantics of events because they include the duration of the event, whereas model fragments are instantaneous descriptions of something that is continuously ongoing. This allows us to mention time in qualitative proportionalities (e.g. as in Figure 2, the time a motion takes is qualitatively inversely proportional to its speed, all else being equal).

While most of the representational machinery is in place for examples such as these, there are invariably gaps. For example, while NextKB currently has some information about snowboarding and snowboards, there is no fact expressing that snowboards are used in snowboarding[3]. The NextKB lexicon contains valence patterns for mapping syntactic patterns to semantic relations, e.g. "on" should get mapped to surfaceObject for snowboarding, but isn't currently. In addition to hand-engineering, we are exploring ways for such gaps to be filled interactively by people using the system (Nakos & Forbus, 2024).

The qualitative models for the two events will be the base and target of an analogical comparison, to construct the mapping that aligns their parameters based on the causal qualitative models. Note that the models do not need to be identical, since analogy is capable of partial matches, but the input and query quantities need to be correctly aligned for the reasoning to be accurate.

6 Extracting cross-situation facts, inputs, and outputs

In addition to extracting the situations to be compared, what is known about their similarities and differences must be extracted. In Example 1, it is the faster movement over ice that is relevant. In Example 2, it is the existence of a period of inactivity plus running versus entirely running that is the key difference. The variety in terms of types of information and

[3] NextKB is derived from OpenCyc, which is an open-source subset of the Cyc KB that contains a small fraction of what Cyc contains. We are slowly filling in, by hand and by ML, such gaps as we need to, in order to continue developing an open-license resource. https://www.qrg.northwestern.edu/nextkb/index.html

Second, identify the quantity type being asked about, based on the scenario and the multiple choice answers A and B.
This part of your response should be of the form
Query: <quantity> <choiceA>
where <quantity> is the type of quantity being asked about.
<choiceA> is 1 if the answer choice A implies that quantity is larger in <descriptionA> than <descriptionB>, and -1 otherwise.
Third, identify a different quantity that varies between <descriptionA> and <descriptionB>, based on the scenario.
If that quantity is larger in <descriptionB>, then use the value 1, and if it is smaller in <descriptionB>,
then use the value -1.
The output for this third part should be of the form
Input: <iquantity> <value>
where <iquantity> is the quantity type just identified and <value> is the value.
You must not say anything else, either before the four lines of output or after them.
You must not provide any other information on the four lines than what was requested.
Do not comment or produce any other output besides the four lines specified here.

Figure 4: Extracting Query and Difference Parameters

how it is expressed in language suggests trying to extract this information via LLM.

Figure 4 illustrates a prompt that attempts to extract such information. The label "Query:" is used to identify the parameter being sought and how to select an answer based on the difference found. That is, the alignment between the output of the DQ analysis (1, 0, or -1) to which answer choice should be made (here, A or B) is handled by the output provided for <ChoiceA>. The other relevant parameter is to be labeled "Input:", and the LLM is tasked with both identifying the parameter and identifying which value should be input to the DQ analysis process (i.e., -1, 0, or 1)[4].

Notice the admonishments against additional output from the LLM. The same training that leads them to be instructable to some degree via natural language also tends to lead to a verbose and obsequious tendency to over-generate. One hopes that as the technology evolves, training regimes are developed that preserve instructability but produce systems that are more suitable to be used as components rather than standalone systems.

For Example 1, Psi4 with this prompting produces

Query: smoothness 1

Input: friction -1

In other words, smoothness being higher means the system should choose A, and that friction is smaller in the second choice (i.e. ice). For Example 2, Psi4 produces

Query: distance run 1

Input: rest duration 1

That is, the answer choice A (John) implies the distance parameter is larger, and the duration of rest is larger in the second situation.

Note that these need to be mapped into the model via CNLU. It may be worth using sequential prompting, so that few-shot training on how to express the quantities in the model built by CNLU's parsing of the first prompt can be provided to tune the second stage to generate more easily usable outputs.

7 Carrying out the qualitative reasoning

If the query and input parameters are correctly identified, then this step is straightforward. Consider the ordinal relationship between the input parameters across the base and target. If the input for base is larger than the target, the DQ value for input is -1, that is, it decreases from base to target. If larger, then the DQ value is 1, and if the same, then DQ = 0. To derive the DQ value for the query parameter, consider the sign of the qualitative proportionality. If positive, then DQ of the query parameter is the same as the input parameter. If negative, then it is the opposite. Note that, with respect to this form of differential qualitative analysis it is symmetric with respect to the causal direction of the qualitative proportionality.

If the query or input parameters are misidentified, then the system should backtrack and explore alternatives, e.g. alternate resolutions for semantic interpretation. The failure modes include not being able to identify the input or query quantity correctly, which would involve re-prompting. Another potential failure mode is mis-alignment in the analogical matching, which should be rare given that the vast majority of QuaRel problems involve pairs of identical events and hence identical instances of causal models.

8 Conclusions and Future Work

This paper explored how a hybrid consisting of an LLM and a symbolic NLU system might be used to support model formulation and interpretation. While this is work in progress, the tests we have made to date are encouraging.

[4] There is an optimization here of ignoring the possibility that the values are the same. That does not happen in QuaRel problems, but does in other tests that involve comparative analysis, such as the Bennett Mechanical Comprehension Test (Klenk et al. 2005).

We see two next steps. The first is to extend the components of the system to handle the full process of solving QuaRel problems, including backtracking as needed to handle the inevitable misinterpretations that occur when natural language is involved. The second step is to extend it to handle the QuaRTz dataset, which is more open-format. In parallel with these efforts, we plan to have Companions use the training sets that come with these datasets as an opportunity to fill in knowledge gaps, by accumulating and generalizing from incidental information provided in the problems.

Finally, one problem with LLMs as a model of statistical commonsense is that, for people, we often have relevant episodic memories that we can draw upon when trying to make sense of a new situation. Such memories do not exist in LLMs, and it would take large-scale learning in a cognitive architecture to explore this prospect.

Ethical Statement

There are no ethical issues.

Acknowledgments

This research was supported by the Office of Naval Research. Useful feedback was provided by Wangcheng Xu and Omar Khater.

References

Allen, J. F. 1994. *Natural Language Understanding. (2nd ed).* Redwood City, CA.: Benjamin/Cummings

Bredeweg, B., Linnebank, F., Bouwer, A., & Liem, J. (2009). Garp3—Workbench for qualitative modelling and simulation. *Ecological Informatics*, 4(5-6), 263-281.

Crouse, M., MFate, C.J., and Forbus, K.D. (2018a). Learning from Unannotated QA Pairs to Analogically Disambiguate and Answer Questions. *Proceedings of AAAI 2018*.

Crouse, M., McFate, C.J., & Forbus, K. (2018b). Learning to Build Qualitative Scenario Models from Natural Langauge. *Proceedings of QR 2018*, Stockholm.

Forbus, K. (1984). Qualitative process theory. *Artificial Intelligence*, 24, 85-168.

Forbus, K. (2019). *Qualitative Representations: How People Reason and Learn about the Continuous World*, MIT Press.

Forbus, K. (2023). Domain Theories for Commonsense Reasoning from Language-Grounded Ontologies. *Proceedings of QR 2023*, Krakow, Poland.

Forbus, K., Carney, K., Sherin, B. and Ureel, L. (2004). VModel: A visual qualitative modeling environment for middle-school students. *Proceedings of the 16th Innovative Applications of Artificial Intelligence Conference*, San Jose, July 2004.

Kamp, H. and Reyle, U. 1993. *From Discourse to Logic: Introduction to Model theoretic Semantics of Natural Language*. Kluwer Academic Dordrecht; Boston.

Klenk, M., Forbus, K., Tomai, E., Kim,H., and Kyckelhahn, B. (2005). Solving Everyday Physical Reasoning Problems by Analogy using Sketches. *Proceedings of 20th National Conference on Artificial Intelligence* (AAAI-05), Pittsburgh, PA.

La Malfa, E. Petrov, A., Frieder, S., Weinhuber, C., Burnell, R., Nazar, R., Cohn, A., Shadbolt, N., & Wooldridge, M. (2025). Language-Models-as-a-Service: Overview of a New Paradigm and its Challenges. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27), 28742-28742

Leelawong, K. & Biswas, G. (2008). Designing learning by teaching agents: The Betty's Brain system. *International Journal of Artificial Intelligence in Education*, 18(3), 181-208.

Nakos, C. & Forbus, K. (2024) Interactively Diagnosing Errors in a Semantic Parser. *Advances in Cognitive Systems Conference*, 2024 https://arxiv.org/abs/2407.06400

Ribeiro, D., Hinrichs, T., Crouse, M., Forbus, K., Chang, M., and Witbrock, M. (2019). Predicting State Changes in Procedural Text using Analogical Question Answering. In *Proceedings of the Seventh Annual Conference on Advances in Cognitive Systems*. Cambridge, MA.

Ribeiro, D. & Forbus, K. (2021). Combining Analogy with Language Models for Knowledge Extraction, *Proceedings of the Third Conference on Automatic Knowledge Base Construction*.

Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S. & Farajtabar, M. (2025) The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity. https://ml-site.cdn-apple.com/papers/the-illusion-of-thinking.pdf

Suzuki, S., & Yoshioka, M. (2024) Preliminary Experiments of Qualitative Reasoning Model Construction Using Large Language Model, *Proceedings of QR2024*, Santiago de Compostela, Spain.

Tafjord, O., Clark, P., Gardner, M., Yih, W. & Sabharwal, A. (2018) QuaRel: A Dataset and Models for Answering Questions about Qualitative Relationships, *AAAI-2018*.

Tafjord, O., Gardner, M., Lin, K. & Clark, P. (2019) QuaRTz: An open-domain dataset of qualitative relationship questions. *EMNLP 2019*.

Tomai, E. and Forbus, K. (2009). EA NLU: Practical Language Understanding for Cognitive Modeling. *Proceedings of the 22nd International Florida Artificial Intelligence Research Society Conference.* Sanibel Island, Florida.