# Causal Knowledge Graphs for Natural Language Understanding

Adi Kalyanpur
Senior Research Scientist,
Elemental Cognition (http://ec.ai)

# Story Cloze Test (Mostafazadeh et al., 2016)

## Narrative comprehension benchmark

***Context***:

Jim got his first credit card in college. He didn't have a job so he bought everything on his card. After he graduated he amounted a $10,000 debt. Jim realized that he was foolish to spend so much money.

***Two alternative endings:***

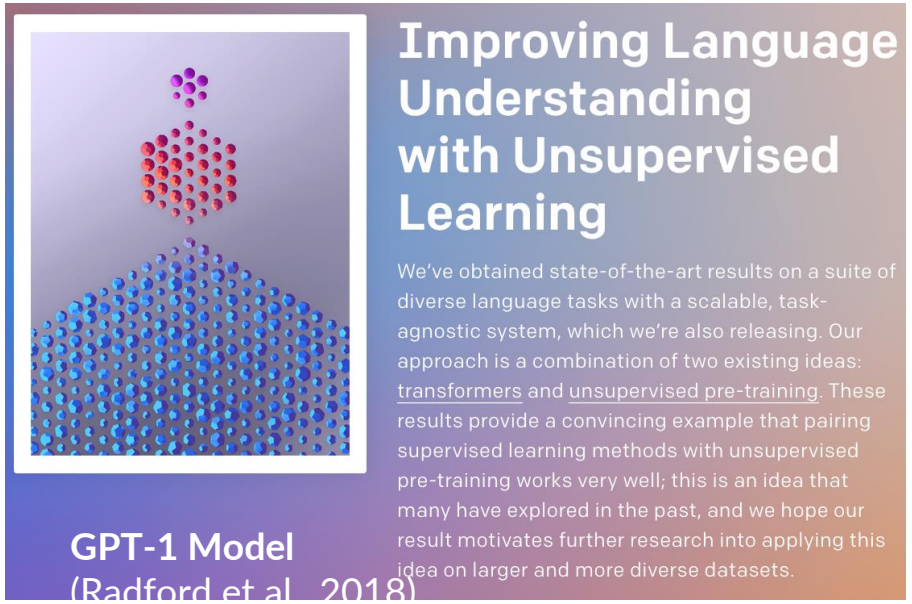
Jim decided to devise a plan for repayment. ✓

Jim decided to open another credit card.

A challenging commonsense reasoning task, where SOTA was ~65% for many months after release of the dataset.

# Things got interesting in 2017/18!

- ***Late 2017-2018:***
  - ***What happened***: The dawn of "Attention is All you need" (Vaswani et al., 2017), introduced **transformers** (non-recursive, attention-based neural models)
  - Large pretrained transformer-based models (BERT, GPT) were *fine-tuned* on downstream NLP tasks, even with little supervised data, and achieved SOTA results!

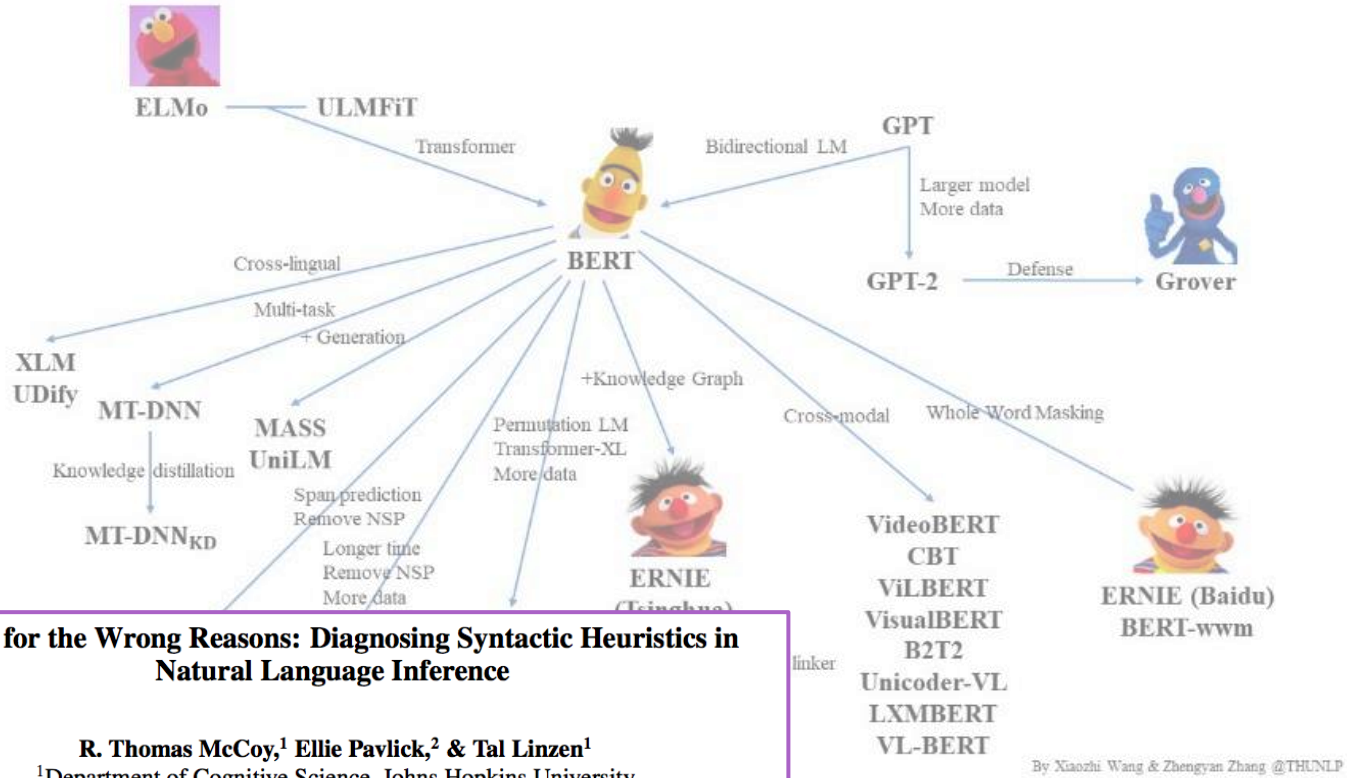



**GPT-1 Model** (Radford et al., 2018)

| DATASET | TASK | SOTA | OURS |
|---|---|---|---|
| ROCStories | Commonsense Reasoning | 77.6 | **86.5** |
| COPA | Commonsense Reasoning | 71.2 | **78.6** |

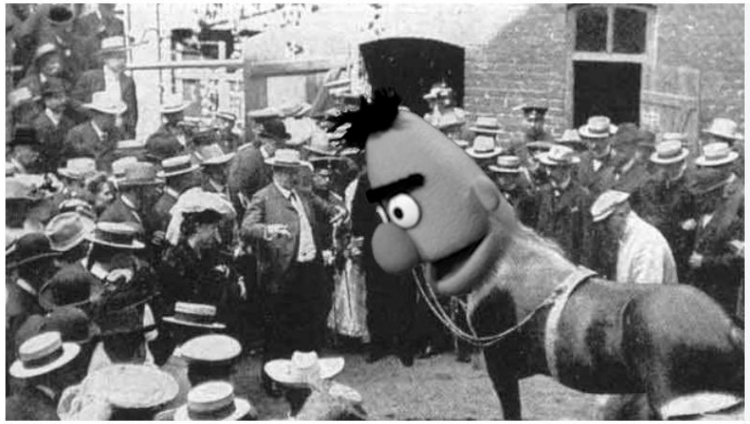
# Last few years…

- Transformer models in full bloom!
- The community has started to think about the weaknesses of E2E deep learning architectures




**Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference**

**R. Thomas McCoy,[1] Ellie Pavlick,[2] & Tal Linzen[1]**
[1]Department of Cognitive Science, Johns Hopkins University
[2]Department of Computer Science, Brown University
tom.mccoy@jhu.edu, ellie_pavlick@brown.edu, tal.linzen@jhu.edu



NLP's Clever Hans Moment has Arrived

26.AUG.2019

Benjamin Heinzerling

It is now almost a cliché to find out that BERT (Devlin et al., 2019) performs "surprisingly well" on whatever dataset you throw at it.

**Test Machine Comprehension, Start by Defining Comprehension**

**Authors:** Jesse Dunietz, Greg Burnham, Akash Bharadwaj, Owen Rambow, Jennifer Chu-Carroll, & David Ferrucci

**Publication venue:** ACL 2020

**But how well do these models understand text?**
**Do they transfer well across domains and tasks?**

Our moonshot at ELEMENTAL cognition

*Machines as thought partners*

We are working on AI systems that **can read, reason and understand text**
by building **rich logical models** of the **meaning** underlying it.
The AI **collaborates** with a human to build a **shared understanding**

# When humans, even young children, read, they make countless implicit commonsense inferences that frame their understanding of the unfolding narrative

Peppa was riding her bike. A car turned in front of her. Peppa turned her bike sharply. She fell off of her bike. Peppa skinned her knee.

While reading, humans construct a coherent representation of what happened and ***why***, combining information from the text with relevant background knowledge

# Humans can construct the causal chain that describes how the sequence of events led to a particular outcome

A car turned in front of Peppa *causes* →

Peppa to turn her bike sharply causes→

*Peppa* fell off of her bike

causes→

Peppa skinned her knee

causes→

(likely) she asks for help!

# Humans can also describe how characters' different states, such as emotions and location, change throughout the story

Peppa was on her bike throughout riding it.

Then after falling, Peppa was on the ground.

Peppa went from feeling (likely) happy to feeling in pain after falling.

Though humans build such mental models of situations with ease (Zwaan et al., 1995), **AI systems** for tasks such as reading comprehension and dialogue **remain far from exhibiting similar commonsense reasoning capabilities**

Why?

- Two major bottlenecks in the AI research

**Not having ways to acquire (often-implicit) commonsense knowledge at scale**

**Not having ways to incorporate knowledge into the state-of-the-art AI systems**

# GLUCOSE: **G**eneraLized and **CO**ntextualized **S**tory **E**xplanations

**A new commonsense reasoning framework for tackling both those bottlenecks at scale**

**(EMNLP 2020 – Honorable Mention Best Paper)**

Authors: Nasrin M, Adi K, Lori M, David B, Lauren B, Or B, Jennifer C

# GLUCOSE Commonsense Reasoning Framework

- Given a short story *S* and a selected sentence *X* in the story, GLUCOSE defines **ten dimensions of commonsense causal explanations** related to X, **inspired by human cognitive psychology**.

| Question | Example answer |
|---|---|
| 1. Did any other **events** directly cause or enable this event? | ***A car turned in front of him*** CAUSED/ENABLED *Gage turned his bike* |
| 2. Did any **emotions or basic human drives** motivate this event? | ***Gage wanted safety*** CAUSED/ENABLED *Gage turned his bike* |
| 3. Did any **location states** make this event possible? | ***Gage was close to a car*** ENABLED *Gage turned his bike away from the car* |
| 4. Did any **possession states** make this event possible? | ***Gage possesses a bike*** ENABLED *Gage turned his bike* |
| 5. Did any **other attributes of some entity** make this event possible? | (No relevant attributes in this example) |

Dimensions 6-10 are duals of 1-5 (e.g. Dim 6 is caused/enabled by X)

# GLUCOSE framework through an Example

*Peppa was riding her bike. A car turned in front of her. Peppa turned her bike sharply She fell off of her bike. Peppa skinned her knee.*

Semi-structured Inference Rule = antecedent *connective* consequent

**Dim #1**

Is there an event that directly causes or enables X?

A car (subject) turned (verb) in front of (preposition) her (object) *Causes/Enables* Peppa (subject) turned (verb) her bike (object)

**Contextualized**: Specific statements exemplify how a general rule could be grounded in a particular context

$\text{Sth}_A$ (subject) turns (verb) in front of (preposition) $\text{Sth}_B$ (that is $\text{Someone}_A$'s vehicle) (object) *Causes/Enables* $\text{Someone}_A$ (subject) turns (verb) $\text{Sth}_B$ (object1) away from (preposition) $\text{Sth}_A$ (object2)

**Dim #2**

Is there an emotion or basic human drive that motivates X?

Peppa (subject) wants (verb) safety (object) *Causes/Enables* Peppa (subject) turned (verb) her bike (object)

**Generalized**: General rules provide general mini-theories about the world!

$\text{Someone}_A$ (subject) wants (verb) safety (object) *Causes/Enables* $\text{Someone}_A$ (subject) moves (verb) away from (preposition) $\text{Something}_A$ (that is dangerous) (object1)

**Dim #3**

Is there a location state that enables X?

Peppa (subject) was (verb) close to (preposition) a car (object) *Enables* Peppa (subject) turned (verb) her bike (object1) away from (preposition) the car (object2)

$\text{Someone}_A$ (subject) is (verb) close to (preposition) $\text{Something}_A$ (object) *Enables* $\text{Someone}_A$ (subject) moves (verb) away from (preposition) $\text{Something}_A$ (object2)

# GLUCOSE framework through an Example

*Peppa was riding her bike. A car turned in front of her. Peppa turned her bike sharply. She fell off of her bike. Peppa skinned her knee.*


Is there a possession state that enables X?

Peppa (subject) possesses (verb) a bike (object) *Enables* Peppa (subject) turned (verb) her bike (object)

Someone$_A$ (subject) possesses (verb) Something$_A$ (object) *Enables* Someone$_A$ (subject) moves (verb) Something$_A$ (object)


Are there any other attributes enabling X?

N/A (the dimension is not applicable for this example)

GLUCOSE is a unique perspective on commonsense reasoning for presenting often-implicit commonsense knowledge in the form of ***semi-structured general*** **inference rules** that are also ***grounded*** in the **context of a specific story**

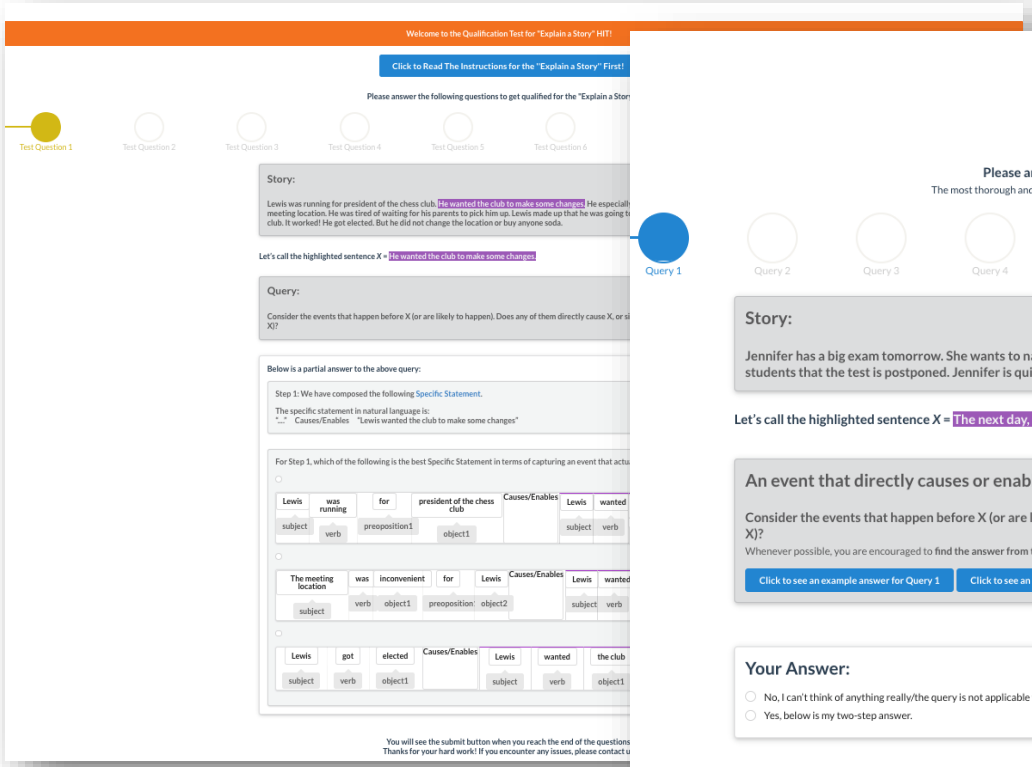
GLUCOSE captures **mini causal theories about the world** focused around **events, states** (location, possession, emotion, etc), **motivations**, and **naive human psychology**.

# How do we address the problem of implicit knowledge acquisition at scale?

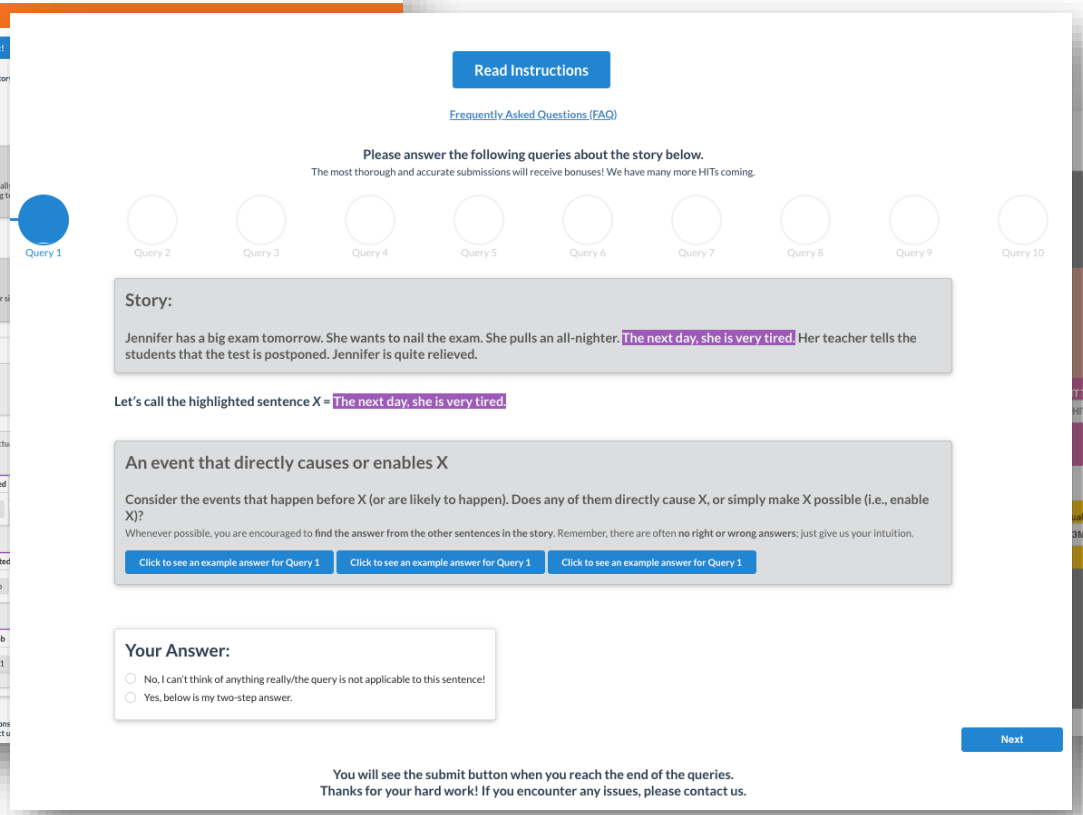
Filling in the GLUCOSE dimensions is **cognitively a complex task** for **lay workers**, since it **requires grasping the concepts of causality and generalization** and to write **semi-structured inference rules**

# An effective multi-stage crowdsourcing platform

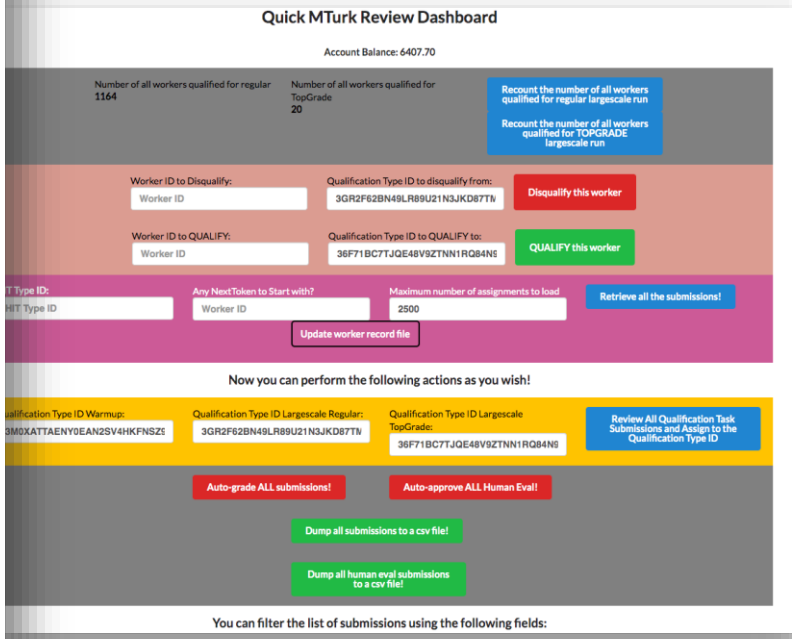
After **many rounds of pilot studies**, we successfully designed an **effective platform** for collecting GLUCOSE data that is **cognitively accessible to laypeople**




GLUCOSE Qualification UI




GLUCOSE Main UI




GLUCOSE Review Dashboard

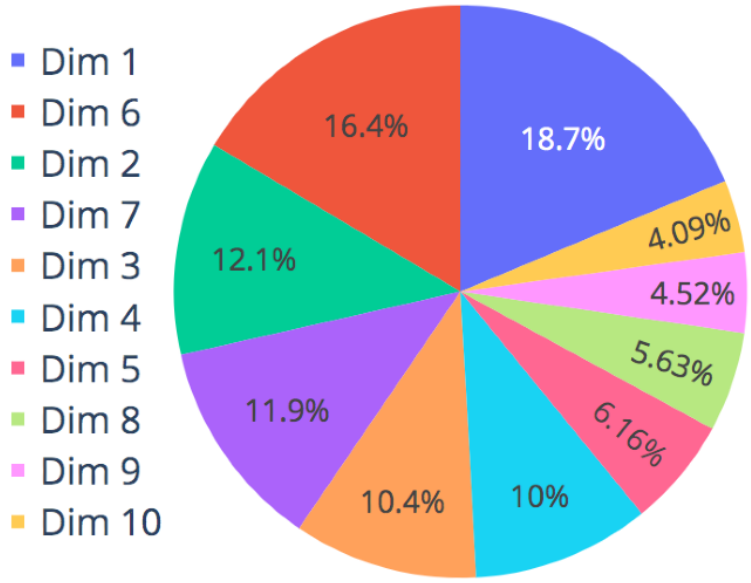
# Statistics and Examples

**Various implicit and script-like mini-theories:**

- Someone_A gives Someone_B Something_A *Results in* Someone_B possess(es) Something_A
- Someone_A is Somewhere_A *Enables* Someone_A forgets Something_A Somewhere_A
- Someone_A is careless *Enables* Someone_A forgets Something_A Somewhere_A
- Someone_A forgets Something_A Somewhere_A *Results in* Something_A is Somewhere_A
- Someone_A feel(s) tired *Enables* Someone_A sleeps
- Someone_A is in bed *Enables* Someone_A sleeps
- Someone_A runs into Someone_B (who Someone_A has not seen for a long time) *Causes* Someone_A feel(s) surprised
- Someone_A asks Someone_B a question *Causes/Enables* Someone_B answers the question

| | |
|---|---|
| # total inference rules | 620K |
| # total unique stories | 4700 |
| # workers participated. | 372 |
| # mins per HIT on avg. | 4.6min |

To our knowledge, GLUCOSE is among the few cognitively-challenging AI tasks to have been successfully crowdsourced



# GLUCOSE captures extensive commonsense knowledge that is unavailable in the existing resources

Ceiling overlap between GLUCOSE and other resources based on best-effort mapping of relations.

| GLUCOSE Dim | 1 | 2 | 5 | 6 | 7 | 10 |
|---|---|---|---|---|---|---|
| **ConceptNet** | 1.2% | 0.3% | 0% | 1.9% | 0% | 0% |
| **ATOMIC** | 7.8% | 1.2% | 2.9% | 5.3% | 1.8% | 4.9% |

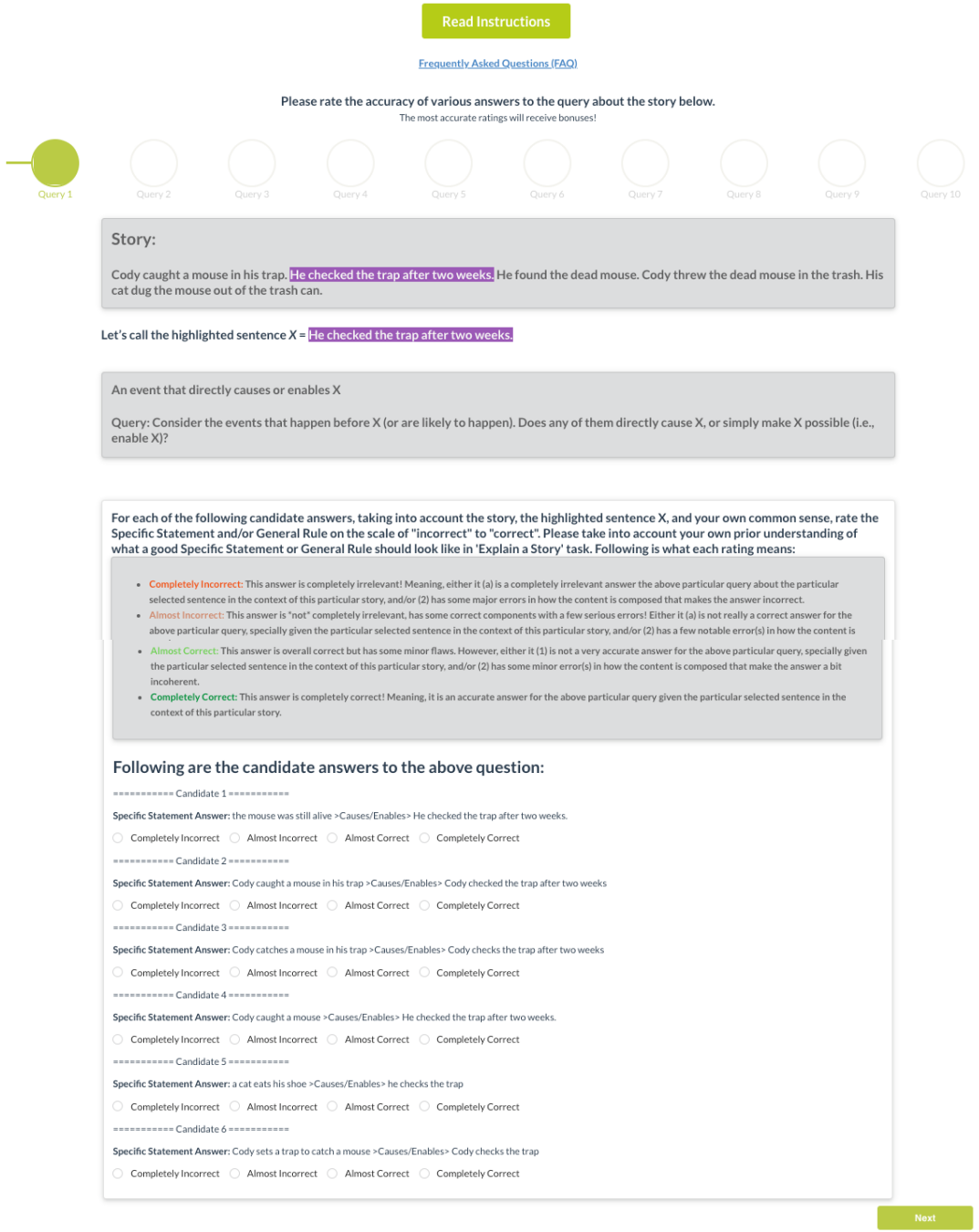
# How to incorporate commonsense knowledge into the state-of-the-art AI systems?

# GLUCOSE Commonsense Reasoning Benchmark

A testbed for evaluating models that can incorporate such commonsense knowledge and show inferential capabilities

- **Task**: Given a story *S*, the sentence *X*, and dimension *d*, predict specific and general rules collected in GLUCOSE
- **Test Set**: We carefully curated a **doubly vetted** test set, based on previously **unseen** stories and on which our **most reliable annotators** had **high** agreement.
  - Our vetting process resulted in a test set of 500 GLUCOSE story/sentence pairs, each with 1-5 dimensions answered.
- **Evaluation Metrics**: Human and Automatic

# We designed a specialized Human Evaluation UI for collecting reliable, reproducible, and calibrated ratings (0-3 Likert scale)



# Notable Models & Results

Baselines

KNN

GPT-2

Avg: g 0.5/3

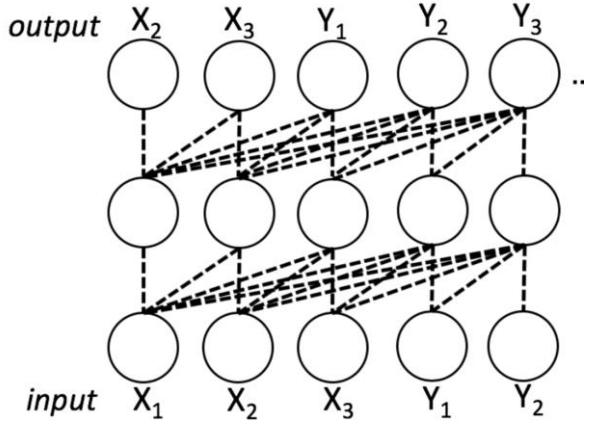
Avg: s 0.8/3

**We show that:**
1) The KNN model performs the worst, highlighting the importance of generalizing beyond the training data.

2) Pre-trained language model perform very poorly at the task and do not show basic commonsense inference

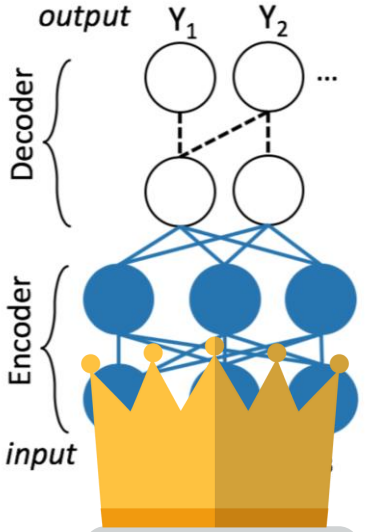
3) When the pre-trained neural models are fine-tuned on the rich GLUCOSE data, they achieve very high performance in making commonsense predictions on ***unseen*** stories.

Trained Models




Full-LM

Avg: s 1.9/3 g 1.7/3




Enc-Dec

Avg: s 2.6/3 g 2.3/3

Human

Avg: s 2.8/3 g 2.6/3

# Example Predictions

*Dimension 3; a location enabling X.*

- Input:
  - *Karen made a pan of lasagna. She brought it to the party. Nobody wanted to eat lasagna. Karen ate it for a week. She became tired of lasagna.*

| | |
|---|---|
| GPT-2 | She was **in front of a TV** >Enables> Karen made a pan of lasagna.<br>N/A |
| Full-LM | Karen is at **home** >Enables> Karen made a pan of lasagna and brought it to the party<br>$Someone_A$ is in $Somewhere_A$ >Enables> $Someone_A$ makes $Something_A$ (that is edible) |
| Enc-Dec | Karen is in **the kitchen** >*Enables*> Karen makes a pan of lasagna<br>$Someone_A$ is in a kitchen >Enables> $Someone_A$ cooks $Something_A$ |
| Human | Karen is in **the kitchen** >*Enables*> Karen made a pan of lasagna<br>$Someone_A$ is in a kitchen >Enables> $Someone_A$ prepares $Something_A$ (that is a dish) |

# Example Predictions

*Dimension 6; an event that X Causes/Enables.*

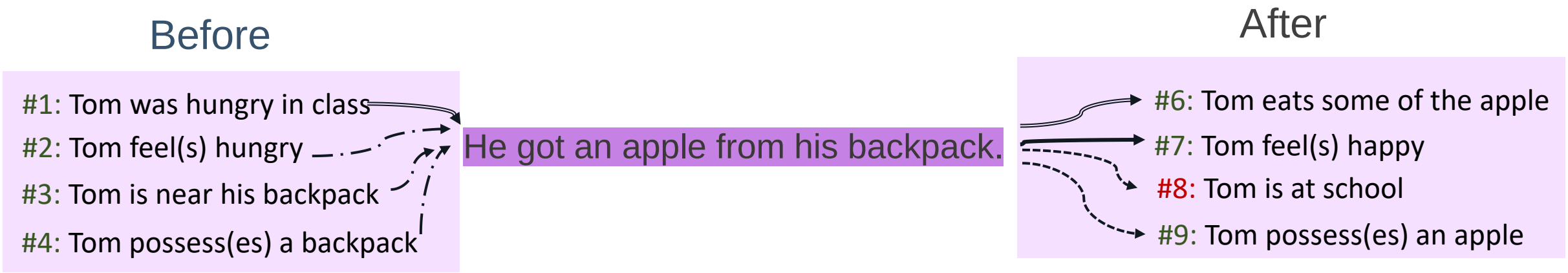
- Input:
  - *Karen made a pan of lasagna. She brought it to the party. Nobody wanted to eat lasagna. Karen ate it for a week. She became tired of lasagna.*

| | |
|---|---|
| Enc-Dec | Karen makes a pan of lasagna >**Causes/Enables**> Karen eats it for a week<br>$Someone_A$ makes $Something_A$ (that is food) >**Causes/Enables**> $Someone_A$ eats $Something_A$ |
| Human | Karen makes a pan of lasagna >**Causes/Enables**> Karen brought it to the party<br>$Someone_A$ prepares $Something_A$ (that is a dish) >**Causes/Enables**> $Someone_A$ takes $Something_A$ to SomethingB (that is an event) |

# Grade:K Story (not from ROC)

Predictions by Enc-Dec model

Tom was hungry in class. He got an apple from his backpack. He ate some of it. It didn't taste good. Tom threw it in the trash. The apple fell behind the trash can. Alice saw it and picked it up. Tom smiled and said "Thank you" to Alice.

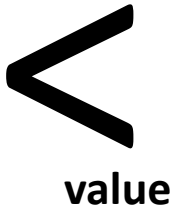


# We verified the following hypothesis

A promising new recipe for giving machines commonsense is to use **high-quality commonsense knowledge** as the **seed data** for training **neural models that have pre-existing lexical and conceptual knowledge**.

Static commonsense knowledge base with GLUCOSE mini-theories authored by humans

**Traditional view of static KB**



**value**

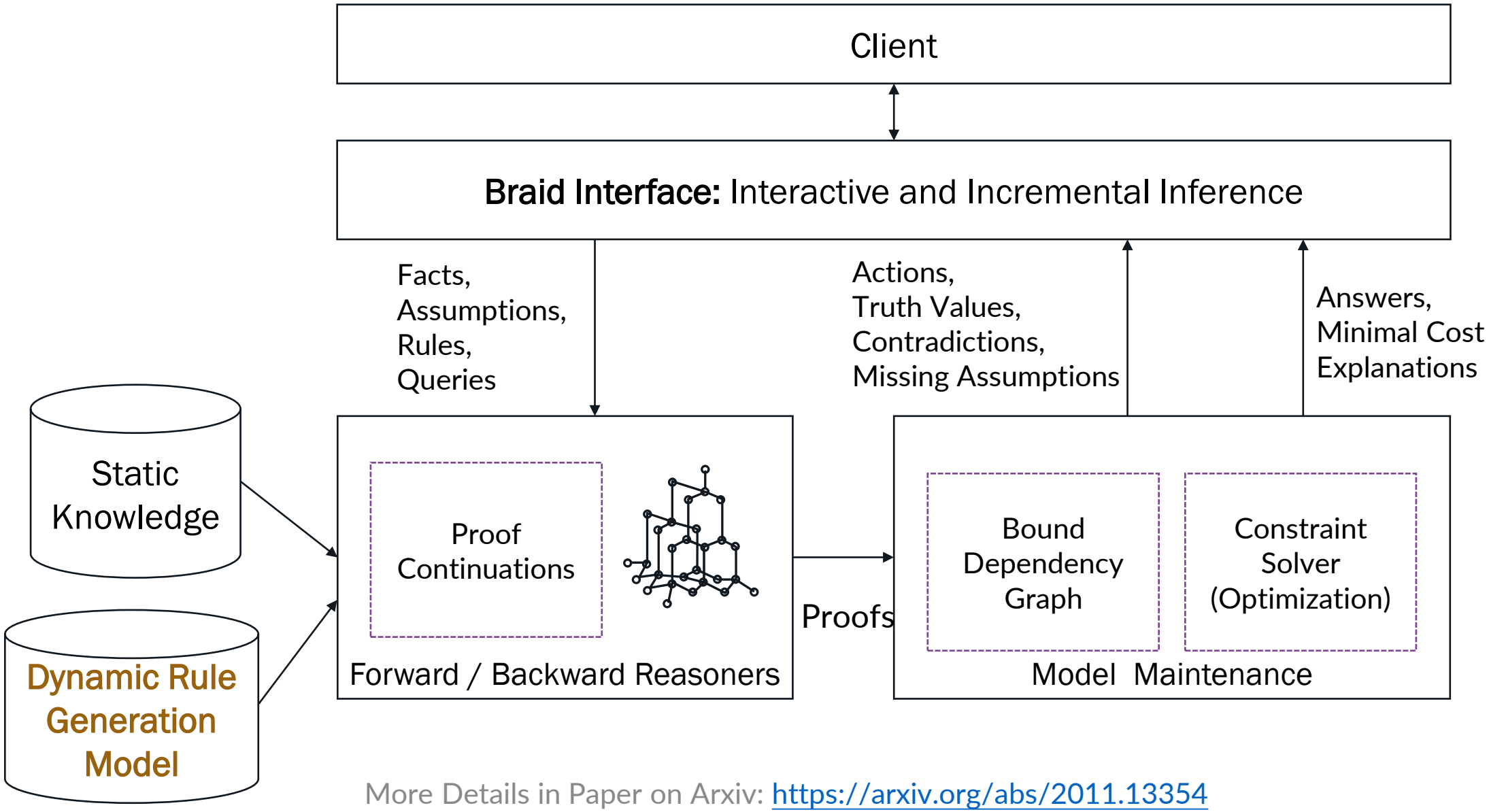
GLUCOSE-Trained model that can generate rules along GLUCOSE dimensions for any novel input

**Dynamic Generative KB**

# Using Casual Knowledge for NLU Reasoning

- At EC, we have built a neuro-symbolic reasoner called **Braid**
- Has foundations in logic programming (Prolog, ASP)
- **Features**
  - Generates logical reasons for the system's answers and ranks explanations based on various aspects (e.g. *plausibility*)
  - Supports integration of statistical functions for unification (fuzzy), **dynamic rule generation**, plausibility-checking etc.
  - Supports *Assumption* based truth-maintenance
  - Does forward & backward chaining, constraint solving, cost-based optimizations
  - Reasons Incrementally and Interactively

# Braid

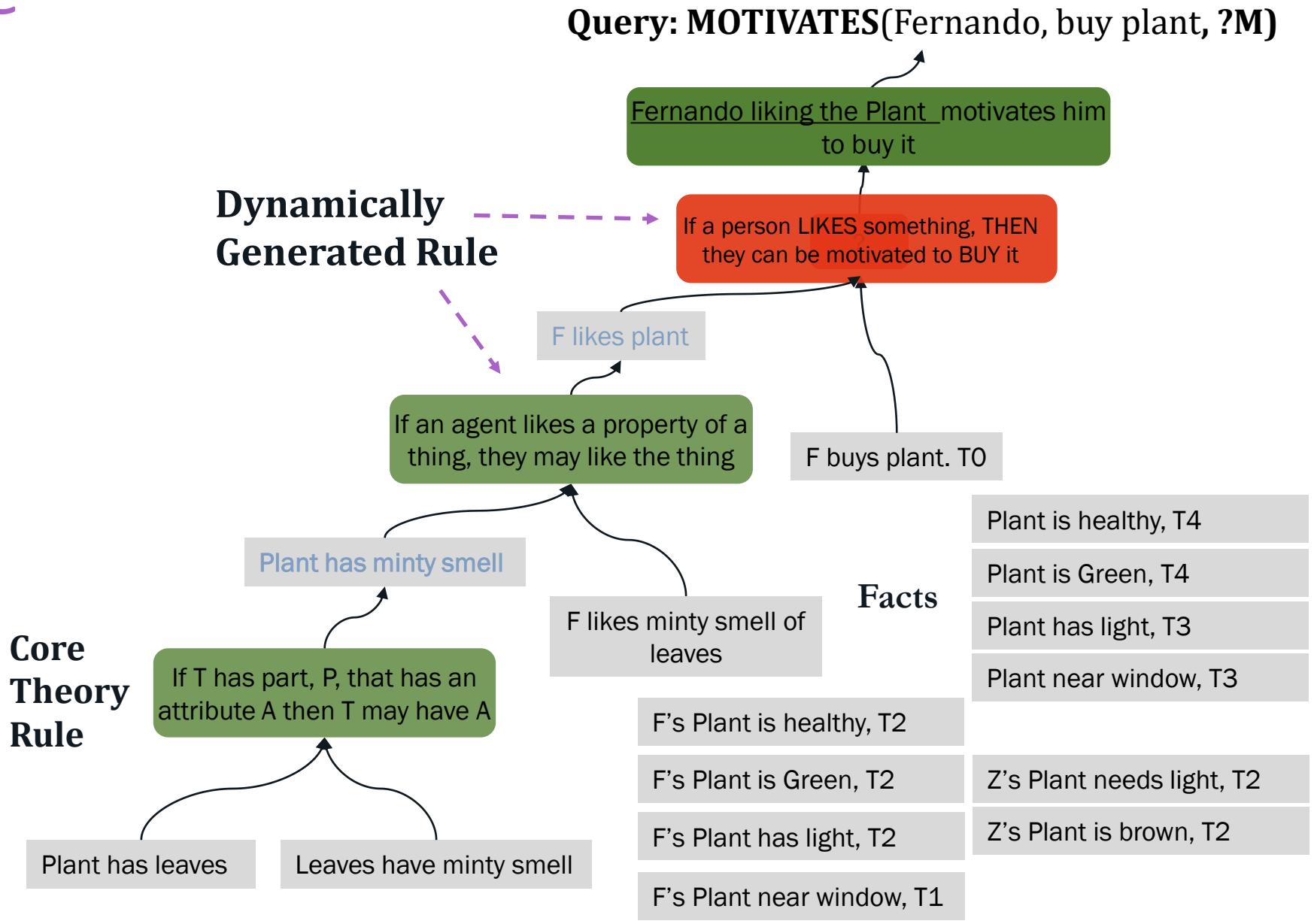



More Details in Paper on Arxiv: https://arxiv.org/abs/2011.13354

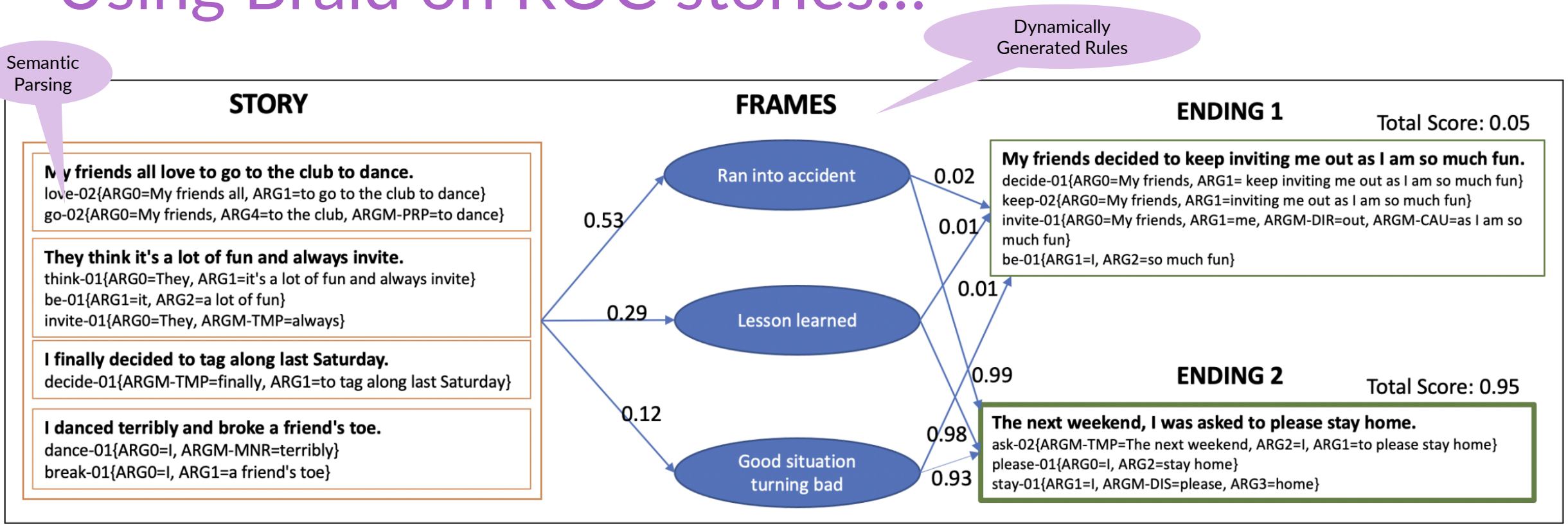
# Braid QA Example

Story: *Fernando goes to a plant sale. He likes the minty smell of leaves. He bought a plant and placed it near a window…*

**QUESTION: Why did Fernando buy a mint plant?**



# Using Braid on ROC stories...




Close to SOTA results *with frame-based explanations*

| Model | Accuracy |
|---|---|
| E2E Neural Baseline | 86.15% |
| Braid-BC: Frame Inf (Text) | 87.17% |
| **Braid-BC: Frame Inf (Sem Parse)** | **87.76%** |
| HintNet (Zhou et al., 2019) | 79.2% |
| GPT2 (Radford et al., 2019) | 86.5% |
| ISCK (Chen et al., 2019) | 87.6% |
| BERT-base + MNLI (Li et al., 2019) | 90.6% |

Table 2: ROC Story Cloze (Spring 2016) Test Results

# To conclude, we show that …

- It is possible to collect **high quality common-sense & causal knowledge from the crowd** with appropriately designed models, tools and UIs
- Fine-tuning pre-trained language models with semi-structured inference rules is an **interesting recipe to do dynamic rule generation in context** (in contrast to a traditional static KB)
- Using **dynamically generated rules for explanation generation in a neuro-symbolic reasoner (Braid)** can alleviate well-known drawbacks of both, traditional KR&R and E2E neural models

# Thanks for listening!

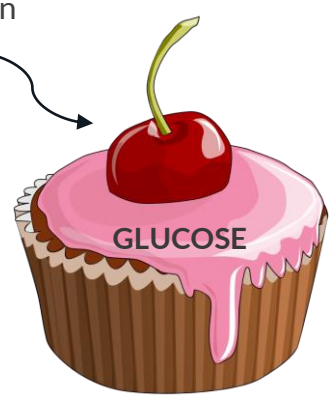
# Automatic Evaluation
## of natural language generations

- A majority of commonsense reasoning frameworks have been in **multiple-choice form, as opposed to natural language generation**, due to **ease of evaluation**
  - Multiple-choice tests are inherently easier to be **gamed**!
- Automatic evaluation for tasks involving natural language generation with diverse possibilities has been a major bottleneck for research
- BLEU's ease of replicability has made it a popular automated metric, but its correlation with human judgement has proven weak in various tasks.

# Automatic Evaluation
## of natural language generations in GLUCOSE

- We found very strong pairwise correlation between human and SacreBLEU corpus-level scores on our test set.
  - Spearman = 0.891, Pearson = 0.855, and Kendall's = 0.705, all with p-value < 0.001.
- This is accomplished through various design choices in GLUCOSE:
  1) GLUCOSE **semi-structured inference rules** are **designed to be evaluable**, where the **structure naturally limits the format** of the generated rules
  2) We curated our test set to eliminate cases with a wide range of correct responses where humans cannot agree, making the **limited number of gold references sufficient for automatic evaluation**
  3) We designed a systematic human evaluation process that can **collect calibrated ratings from judges** who are well educated about what constitutes a correct GLUCOSE rule.

Strong correlation between human and automatic metric!!




GLUCOSE task has a systematic evaluation that is fast and easily replicable!