



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.

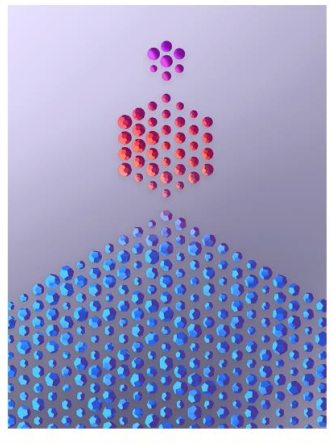


Jim decided to open another credit card.

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!



Improving Language Understanding with Unsupervised Learning

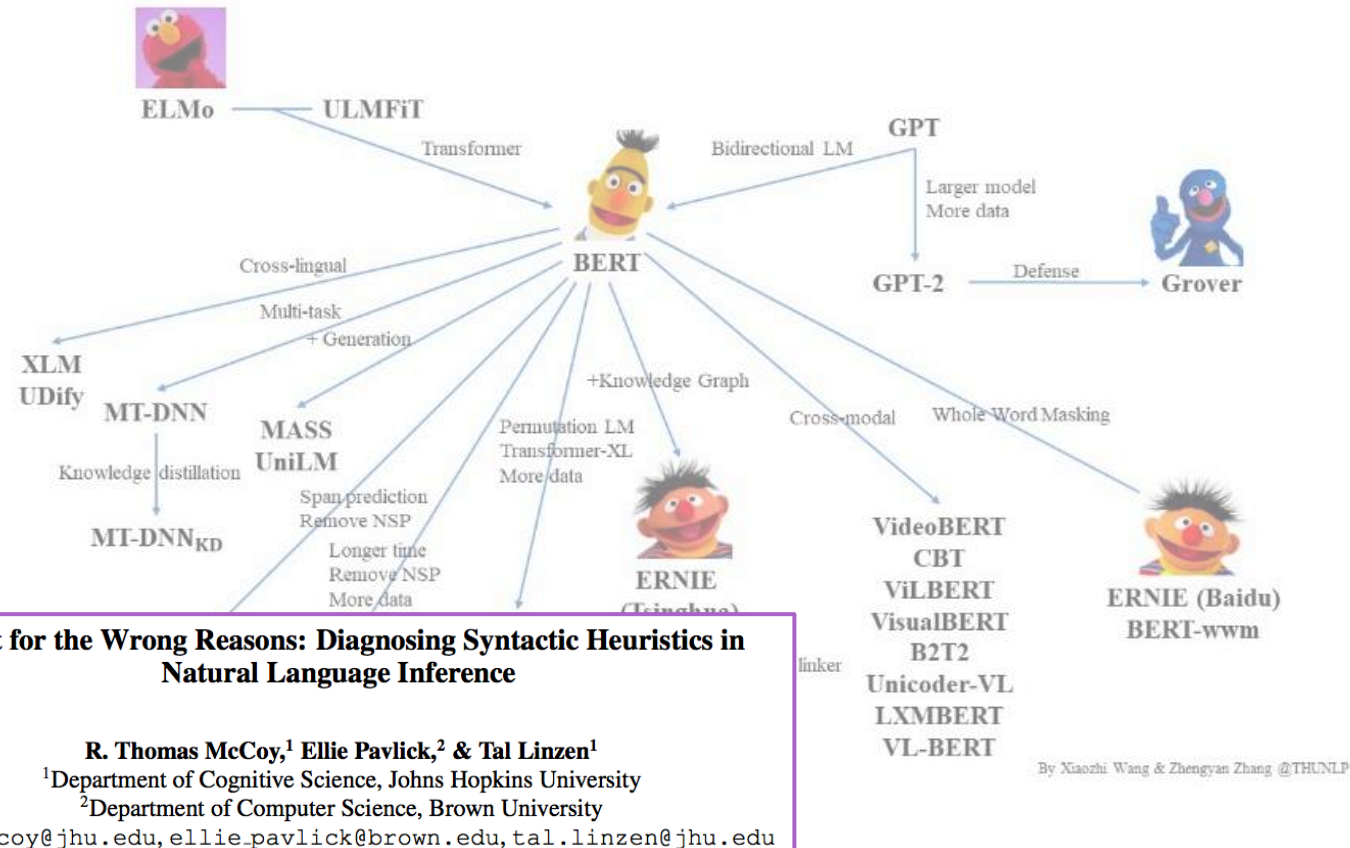
We've obtained state-of-the-art results on a suite of diverse language tasks with a scalable, task-agnostic system, which we're also releasing. Our approach is a combination of two existing ideas: transformers and unsupervised pre-training. These results provide a convincing example that pairing supervised learning methods with unsupervised pre-training works very well; this is an idea that many have explored in the past, and we hope our result motivates further research into applying this idea on larger and more diverse datasets.

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



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: Generalized and Contextualized Story Explanations

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



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

(EMNLP 2020 – Honorable Mention Best Paper)

GLUCOSE Commonsense Reasoning Framework

- Given a short story S and a selected sentence \underline{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?	<i>A car turned in front of him</i> CAUSED/ENABLED Gage turned his bike
2. Did any emotions or basic human drives motivate this event?	<i>Gage wanted safety</i> CAUSED/ENABLED Gage turned his bike
3. Did any location states make this event possible?	<i>Gage was close to a car</i> ENABLED Gage turned his bike away from the car
4. Did any possession states make this event possible?	<i>Gage possesses a bike</i> 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

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

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

Is there an event that directly causes or enables X?

Sth_A turns in front of Sth_B (that is Someone_A's vehicle) *Causes/Enables* Someone_A turns Sth_B away from Sth_A
subject verb preposition object subject verb object1 preposition object2

Dim
#2

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

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

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

Someone_A wants safety *Causes/Enables* Someone_A moves away from Something_A (that is dangerous)
subject verb object subject verb preposition object1

Dim
#3

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

Is there a location state that enables X?

Someone_A is close to Something_A *Enables* Someone_A moves away from Something_A
subject verb preposition object subject verb preposition 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 possesses a bike *Enables* Peppa turned her bike
subject verb object subject verb object

Someone_A possesses Something_A *Enables* Someone_A moves Something_A
subject verb object subject verb 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

Welcome to the Qualification Test for "Explain a Story" HIT!

Click to Read The Instructions for the "Explain a Story" First!

Please answer the following questions to get qualified for the "Explain a Story" HIT!

Test Question 1

Test Question 2

Test Question 3

Test Question 4

Test Question 5

Test Question 6

Story:

Lewis was running for president of the chess club. He wanted the club to make some changes. He especially wanted the club to make some changes. He was tired of waiting for his parents to pick him up. Lewis made up that he was going to club. It worked! He got elected. But he did not change the location or buy anyone soda.

Let's call the highlighted sentence X - He wanted the club to make some changes.

Query:

Consider the events that happen before X (or are likely to happen). Does any of them directly cause X, or simply make X possible (i.e., enable X)?

Below is a partial answer to the above query:

Step 1: We have composed the following Specific Statement.

The specific statement in natural language is: "... Causes/Enables "Lewis wanted the club to make some changes"

For Step 1, which of the following is the best Specific Statement in terms of capturing an event that actually causes or enables X?

Lewis

was

running

for

president of the chess club

Causes/Enables

Lewis

wanted

subject

verb

preposition1

object1

subject

verb

The meeting location

was

inconvenient

for

Lewis

Causes/Enables

Lewis

wanted

subject

verb

object1

preposition:

object2

subject

verb

Lewis

got

elected

Causes/Enables

Lewis

wanted

the club

subject

verb

object1

subject

verb

You will see the submit button when you reach the end of the questions. Thanks for your hard work! If you encounter any issues, please contact us.

Read Instructions

Frequently Asked Questions (FAQ)

Please answer the following queries about the story below.

The most thorough and accurate submissions will receive bonuses! We have many more HITs coming.

Query 1

Query 2

Query 3

Query 4

Query 5

Query 6

Query 7

Query 8

Query 9

Query 10

Story:

Jennifer has a big exam tomorrow. She wants to nail the exam. She pulls an all-nighter. The next day, she is very tired. Her teacher tells the students that the test is postponed. Jennifer is quite relieved.

Let's call the highlighted sentence X - The next day, she is very tired.

An event that directly causes or enables X

Consider the events that happen before X (or are likely to happen). Does any of them directly cause X, or simply make X possible (i.e., enable X)?

Whenever possible, you are encouraged to find the answer from the other sentences in the story. Remember, there are often no right or wrong answers; just give us your intuition.

Click to see an example answer for Query 1

Click to see an example answer for Query 1

Click to see an example answer for Query 1

Your Answer:

No, I can't think of anything really/the query is not applicable to this sentence!

Yes, below is my two-step answer.

Next

You will see the submit button when you reach the end of the queries. Thanks for your hard work! If you encounter any issues, please contact us.

Quick MTurk Review Dashboard

Account Balance: 6407.70

Number of all workers qualified for regular 1164

Number of all workers qualified for TopGrade 20

Recount the number of all workers qualified for regular largescale run

Recount the number of all workers qualified for TOPGRADE largescale run

Worker ID to Disqualify:

Worker ID

Qualification Type ID to disqualify from:

3GR2F62BN49LR89U21N3JKD87TN

Disqualify this worker

Worker ID to QUALIFY:

Worker ID

Qualification Type ID to QUALIFY to:

36F71BC77JQE48V9ZTN1RQB4N6

QUALIFY this worker

Worker ID:

Worker ID

Any Next Token to Start with?

Maximum number of assignments to load

Retrieve all the submissions!

Update worker record file

Now you can perform the following actions as you wish!

Qualification Type ID Warmup:

Qualification Type ID Largescale Regular:

Qualification Type ID Largescale TopGrade:

Review All Qualification Task Submissions and Assign to the Qualification Type ID

Auto-grade ALL submissions!

Auto-approve ALL Human Eval!

Dump all submissions to a csv file!

Dump all human eval submissions to a csv file!

You can filter the list of submissions using the following fields:

GLUCOSE Review Dashboard

GLUCOSE Qualification UI

GLUCOSE Main UI

15

ToC

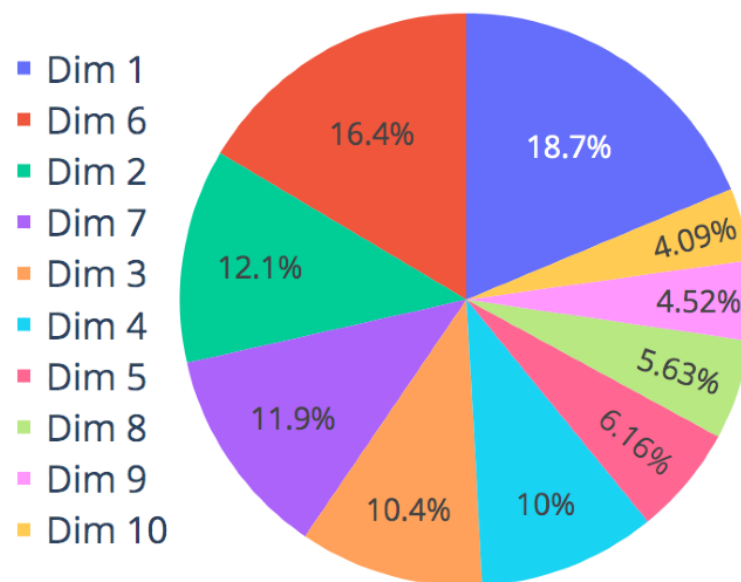
Statistics and Examples

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 Dim1		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 \underline{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)

[Read Instructions](#)
[Frequently Asked Questions \(FAQ\)](#)

Please rate the accuracy of various answers to the query about the story below.
The most accurate ratings will receive bonuses!

Query 1

Query 2

Query 3

Query 4

Query 5

Query 6

Query 7

Query 8

Query 9

Query 10

Story:

Cody caught a mouse in his trap. **He checked the trap after two weeks.** He found the dead mouse. Cody threw the dead mouse in the trash. His cat dug the mouse out of the trash can.

Let's call the highlighted sentence X = **He checked the trap after two weeks.**

An event that directly causes or enables X

Query: Consider the events that happen before X (or are likely to happen). Does any of them directly cause X, or simply make X possible (i.e., enable X)?

For each of the following candidate answers, taking into account the story, the highlighted sentence X, and your own common sense, rate the Specific Statement and/or General Rule on the scale of "incorrect" to "correct". Please take into account your own prior understanding of what a good Specific Statement or General Rule should look like in 'Explain a Story' task. Following is what each rating means:

- Completely Incorrect:** This answer is completely irrelevant! Meaning, either it (a) is a completely irrelevant answer the above particular query about the particular selected sentence in the context of this particular story, and/or (2) has some major errors in how the content is composed that makes the answer incorrect.
- Almost Incorrect:** This answer is "not" completely irrelevant, has some correct components with a few serious errors! Either it (a) is not really a correct answer for the above particular query, specially given the particular selected sentence in the context of this particular story, and/or (2) has a few notable error(s) in how the content is composed.
- Almost Correct:** This answer is overall correct but has some minor flaws. However, either it (1) is not a very accurate answer for the above particular query, specially given the particular selected sentence in the context of this particular story, and/or (2) has some minor error(s) in how the content is composed that make the answer a bit incoherent.
- Completely Correct:** This answer is completely correct! Meaning, it is an accurate answer for the above particular query given the particular selected sentence in the context of this particular story.

Following are the candidate answers to the above question:

***** Candidate 1 *****

Specific Statement Answer: the mouse was still alive >Causes/Enables> He checked the trap after two weeks.

☐ Completely Incorrect ☐ Almost Incorrect ☐ Almost Correct ☐ Completely Correct

***** Candidate 2 *****

Specific Statement Answer: Cody caught a mouse in his trap >Causes/Enables> Cody checked the trap after two weeks

☐ Completely Incorrect ☐ Almost Incorrect ☐ Almost Correct ☐ Completely Correct

***** Candidate 3 *****

Specific Statement Answer: Cody catches a mouse in his trap >Causes/Enables> Cody checks the trap after two weeks

☐ Completely Incorrect ☐ Almost Incorrect ☐ Almost Correct ☐ Completely Correct

***** Candidate 4 *****

Specific Statement Answer: Cody caught a mouse >Causes/Enables> He checked the trap after two weeks.

☐ Completely Incorrect ☐ Almost Incorrect ☐ Almost Correct ☐ Completely Correct

***** Candidate 5 *****

Specific Statement Answer: a cat eats his shoe >Causes/Enables> he checks the trap

☐ Completely Incorrect ☐ Almost Incorrect ☐ Almost Correct ☐ Completely Correct

***** Candidate 6 *****

Specific Statement Answer: Cody sets a trap to catch a mouse >Causes/Enables> Cody checks the trap

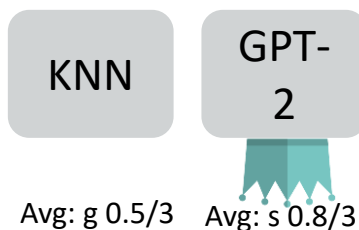
☐ Completely Incorrect ☐ Almost Incorrect ☐ Almost Correct ☐ Completely Correct

[Next](#)

You will see the submit button when you reach the end of the queries.
Thanks for your hard work! If you encounter any issues, please contact us.

Notable Models & Results

Baselines



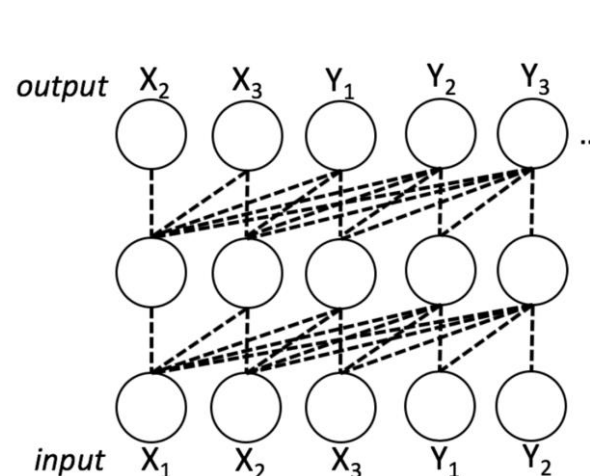
We show that:

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

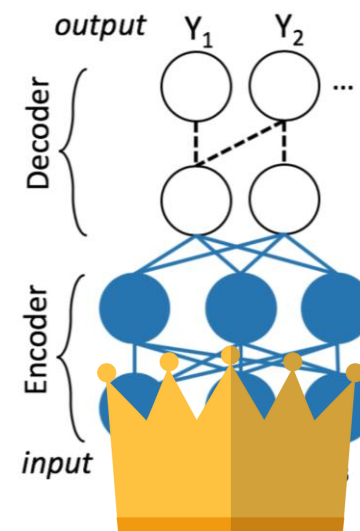
3) When the pre-trained neural models are fine-tuned on the rich GLUCOSE data, they achieve very high performance in making commonsense predictions on **unseen** stories.

Trained Models



Full-LM

Avg: s 1.9/3 g 1.7/3



Enc-Dec

Avg: s 2.6/3 g 2.3/3



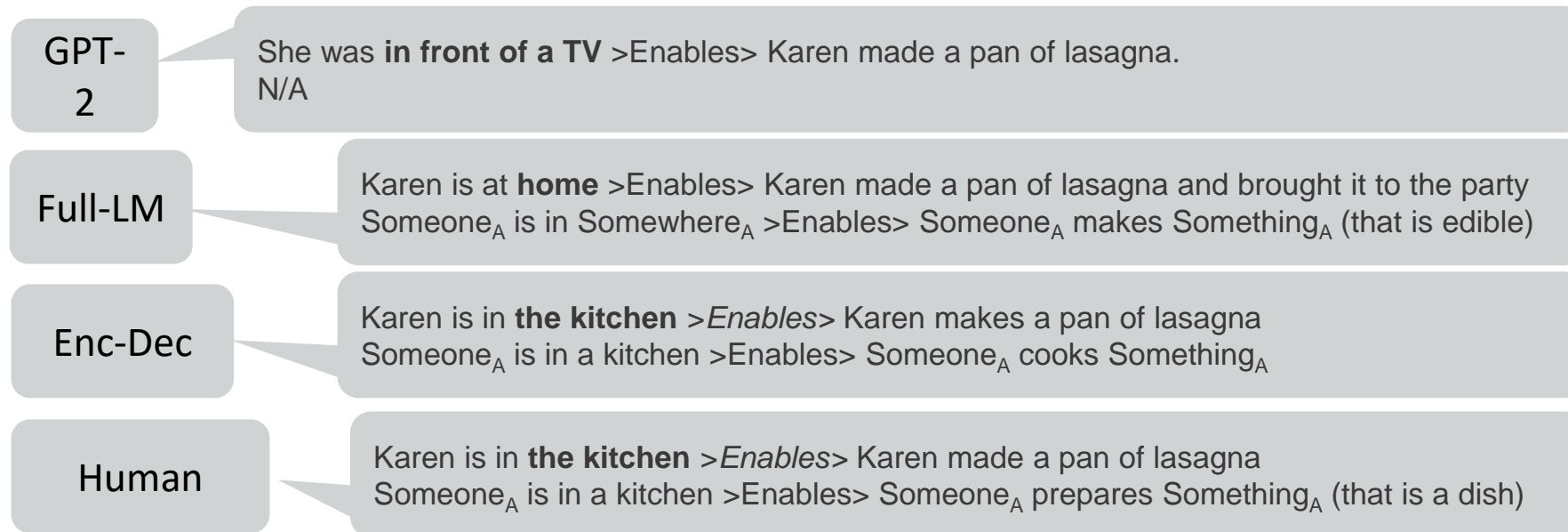
Human

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.*



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
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
Someone_A prepares Something_A (that is a dish) >**Causes/Enables**> Someone_A takes
Something_A to Something_B (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.

Before

- #1: Tom was hungry in class
- #2: Tom feel(s) hungry
- #3: Tom is near his backpack
- #4: Tom possess(es) a backpack

He got an apple from his backpack.

After

- #6: Tom eats some of the apple
- #7: Tom feel(s) happy
- #8: Tom is at school
- #9: Tom possess(es) an apple

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

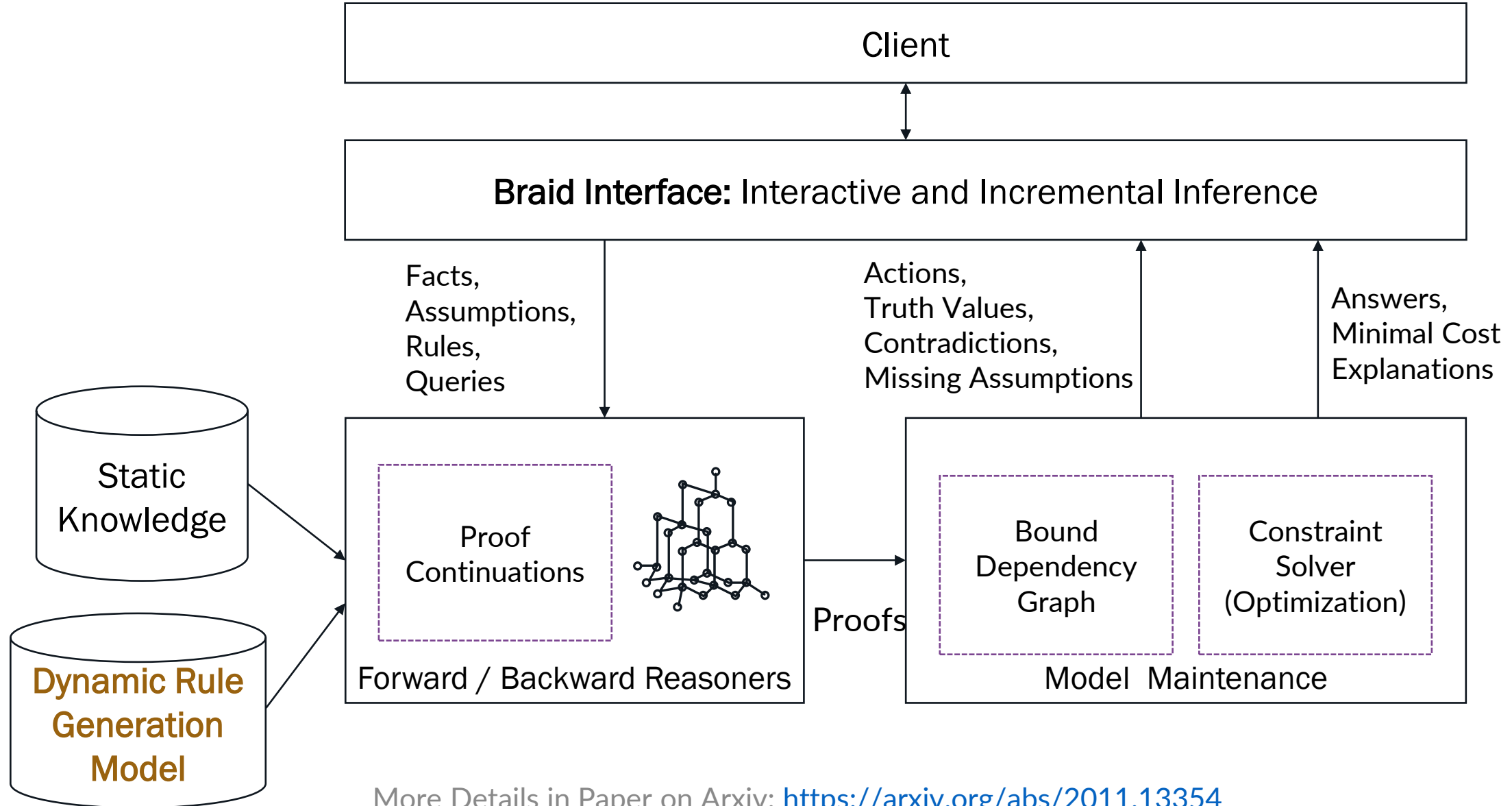


GLUCOSE-Trained model
that can generate rules along
GLUCOSE dimensions for
any novel input

Dynamic Generative KB

Using Casual Knowledge for NLU Reasoning

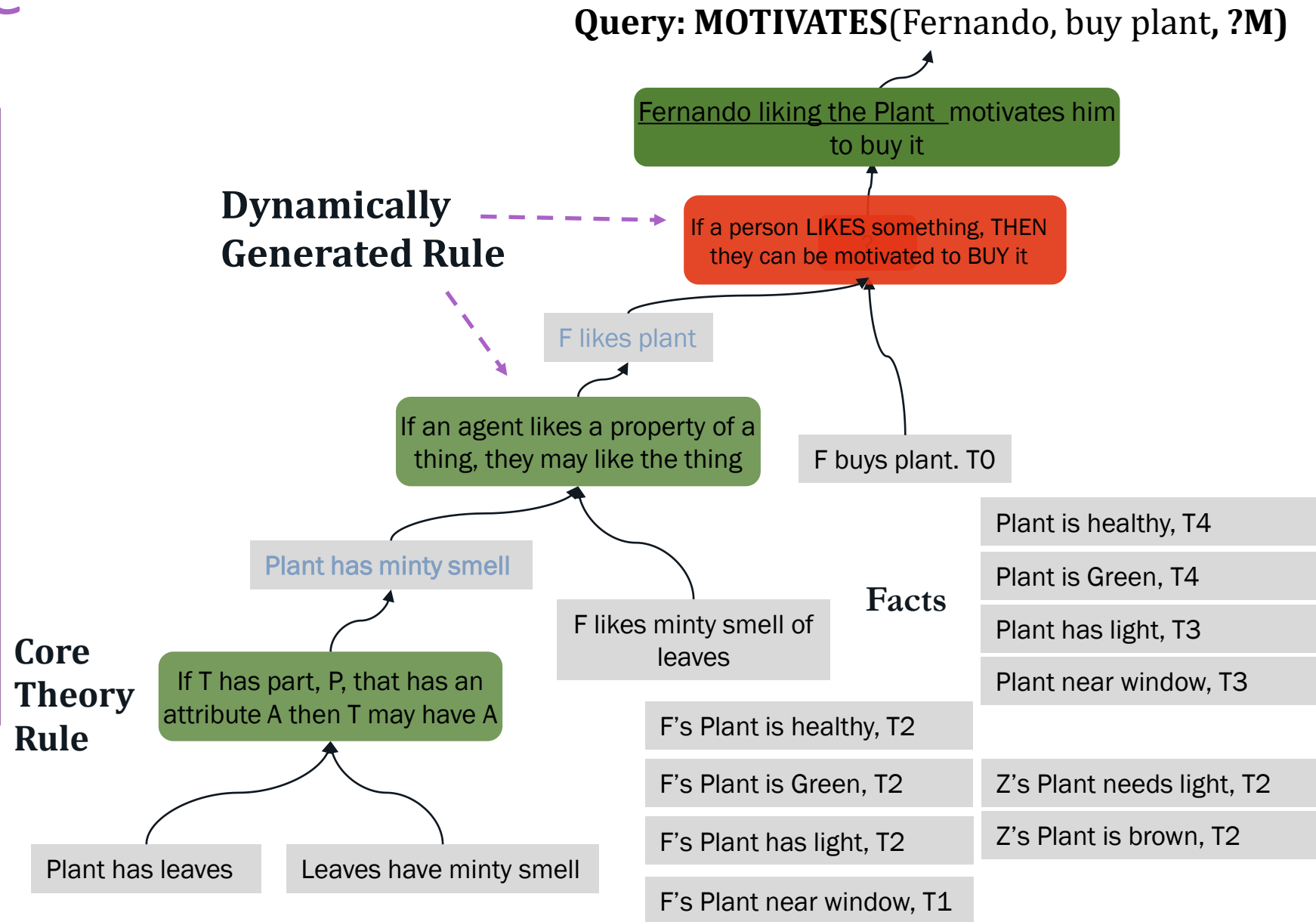
- At EC, we have built a neuro-symbolic reasoner called **Braid**
- Has foundations in logic programming (Prolog, ASP)
- **Features**
 - Generates logical reasons for the system's answers and ranks explanations based on various aspects (e.g. *plausibility*)
 - Supports integration of statistical functions 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 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...

Semantic
Parsing

STORY

FRAMES

Dynamically
Generated Rules

ENDING 1

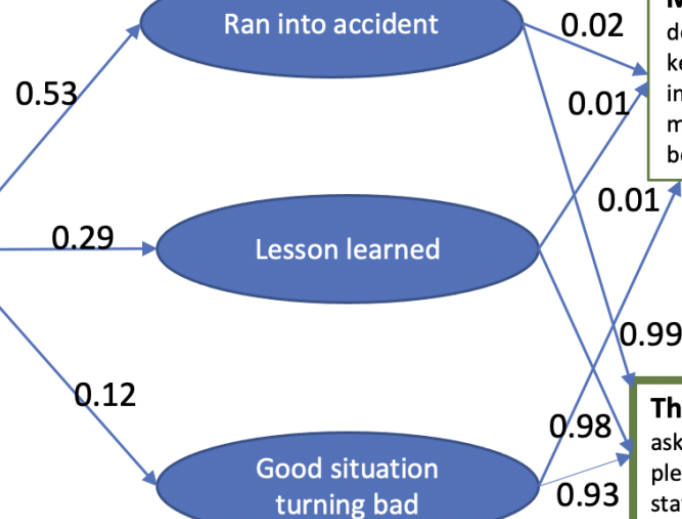
Total Score: 0.05

My friends decided to keep inviting me out as I am so much fun.
 decide-01{ARG0=My friends, ARG1= keep inviting me out as I am so much fun}
 keep-02{ARG0=My friends, ARG1=inviting me out as I am so much fun}
 invite-01{ARG0=My friends, ARG1=me, ARGM-DIR=out, ARGM-CAU=as I am so much fun}
 be-01{ARG1=I, ARG2=so much fun}

ENDING 2

Total Score: 0.95

The next weekend, I was asked to please stay home.
 ask-02{ARGM-TMP=The next weekend, ARG2=I, ARG1=to please stay home}
 please-01{ARG0=I, ARG2=stay home}
 stay-01{ARG1=I, ARGM-DIS=please, ARG3=home}



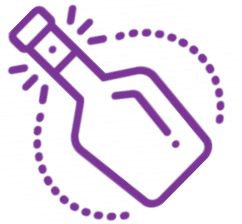
Close to SOTA results with *frame-based explanations*

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

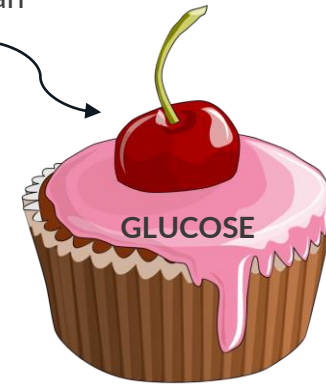
- A majority of commonsense reasoning frameworks have been in **multiple-choice form, as opposed to natural language generation**, due to **ease of evaluation**
 - Multiple-choice tests are inherently easier to be **gamed!**
- Automatic evaluation for tasks involving natural language generation with 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!