Symbolic Knowledge Distillation

From General Language Models to Commonsense Models

— NAACL 2022 —

Peter West

Chandra Bhagavatula

Jack Hessel

Jena Hwang

Liwei Jiang

Ronan Le Bras

Ximing Lu

Sean Welleck

Yejin Choi

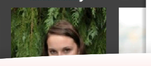
Language models != knowledge models

ATOMIC: Ar
Com
for If-Th

Maarten Sap

Ronan LeBras

Emily Allaway



C
Bha

(COMET-) ATOMIC$^{20}_{20}$:

On Symbolic and Neural Commonsense Knowledge Graphs

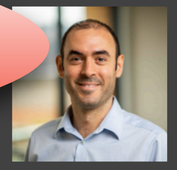
— wait, doesn't GPT-3 know everything? —

AAAI 2021

Jena Hwang

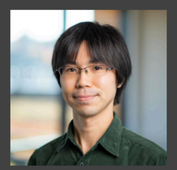
Chandra Bhagavatula

Ronan Le Bras



Jeff Da

Keisuke Sakaguchi



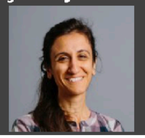
Antoine Bosseult

Me

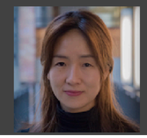
Transformers for
raph Construction

haitanya
Malaviya

Asli Çelikyilmaz



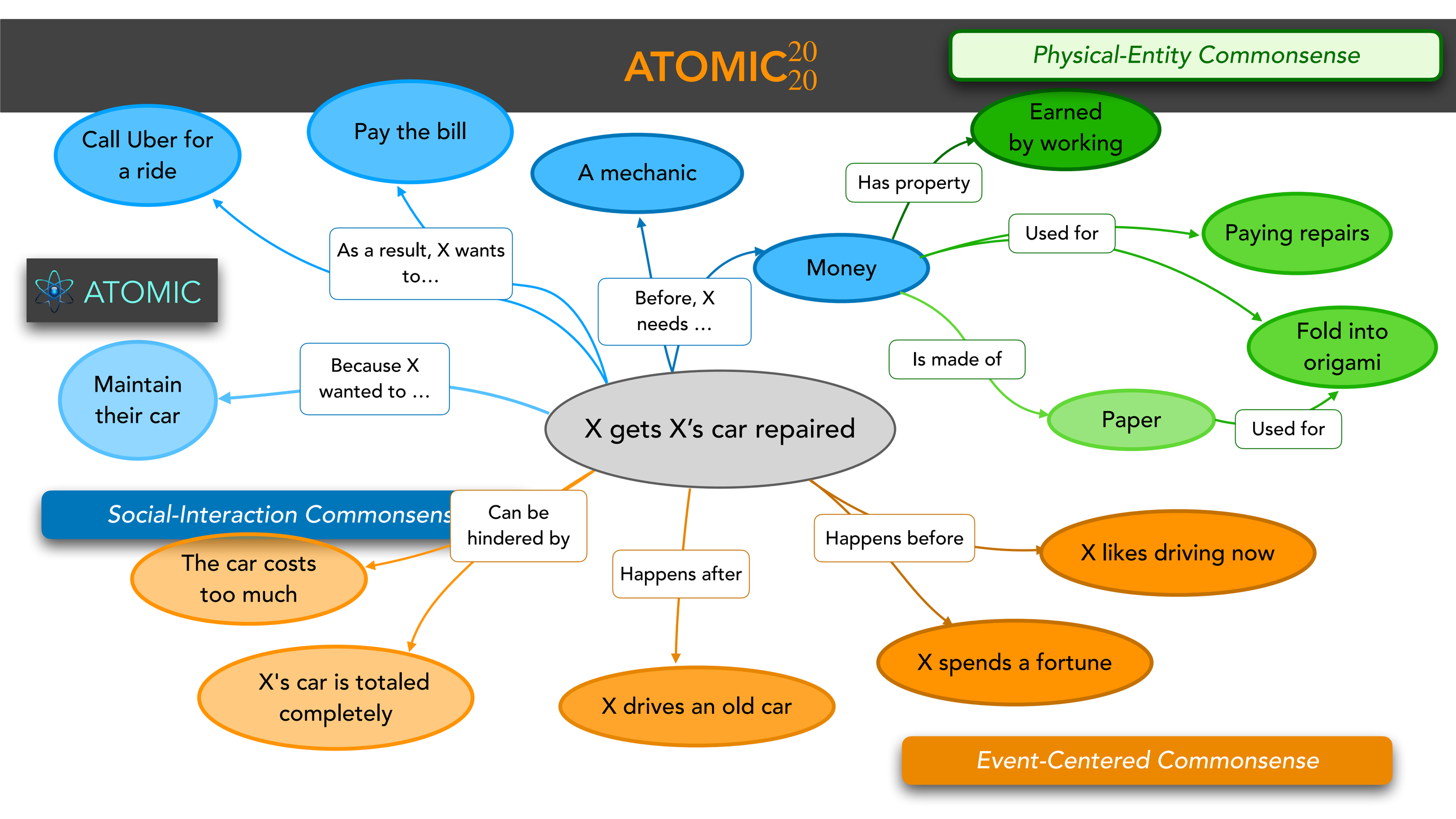
Me



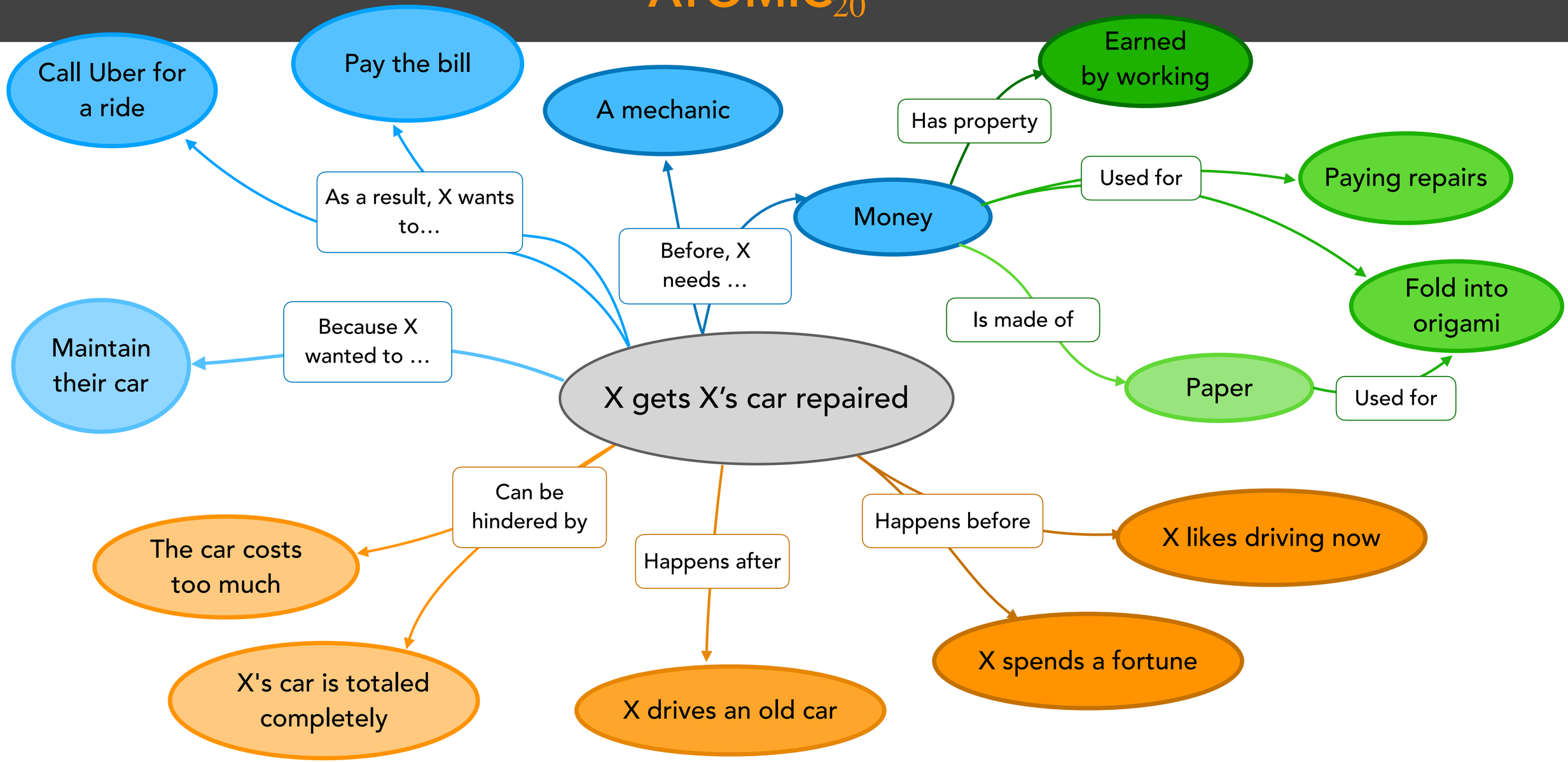
Fully crowdsourced by humans

Symbolic commonsense knowledge graph

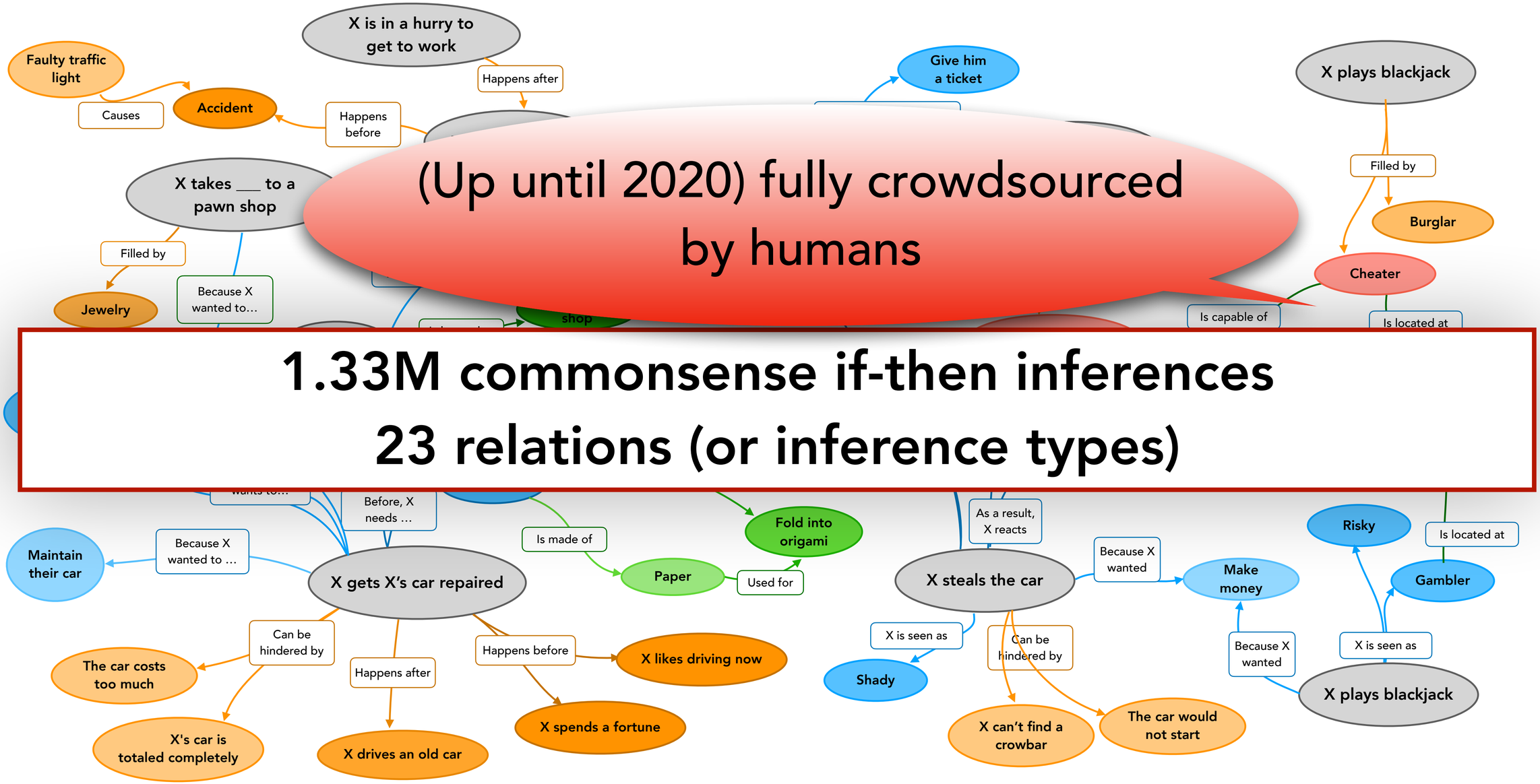
Neural commonsense model


ATOMIC 2020
Physical-Entity Commonsense
Call Uber for a ride
Pay the bill
A mechanic
Earned by working
Has property
As a result, X wants to…
Before, X needs …
Money
Used for
Paying repairs
ATOMIC
Fold into origami
Maintain their car
Because X wanted to …
Is made of
X gets X's car repaired
Paper
Used for
Social-Interaction Commonsens
Can be hindered by
Happens before
X likes driving now
The car costs too much
Happens after
X spends a fortune
X's car is totaled completely
X drives an old car
Event-Centered Commonsense

ATOMIC 2020

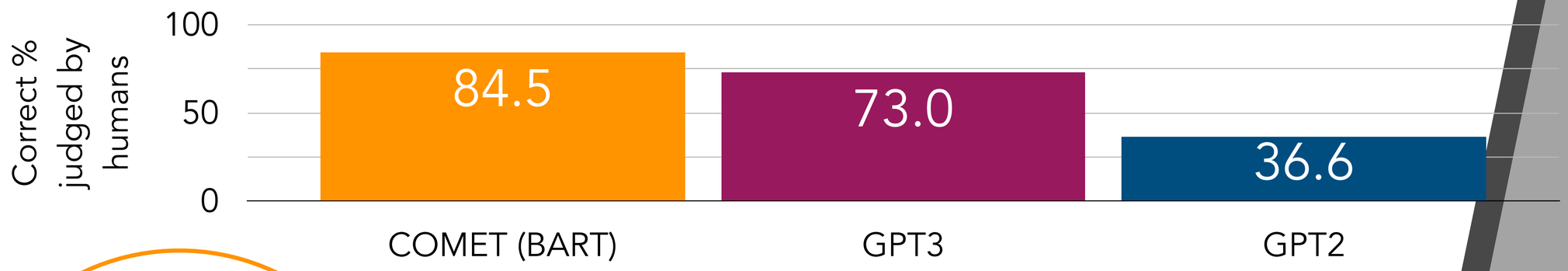

X gets X's car repaired
As a result, X wants to…
Call Uber for a ride
Pay the bill
Before, X needs …
A mechanic
Because X wanted to …
Maintain their car
Money
Has property
Earned by working
Used for
Paying repairs
Is made of
Paper
Used for
Fold into origami
Can be hindered by
The car costs too much
X's car is totaled completely
Happens after
X drives an old car
Happens before
X likes driving now
X spends a fortune

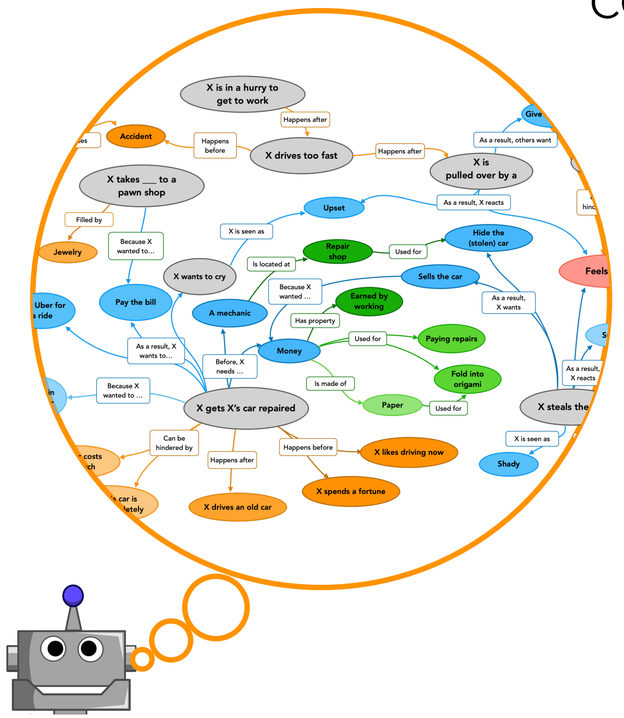
ATOMIC$^{20}_{20}$


X is in a hurry to get to work
Faulty traffic light
Causes
Accident
Happens before
Happens after
Give him a ticket
X takes ___ to a pawn shop
Filled by
Jewelry
Because X wanted to…
X plays blackjack
Filled by
Burglar
Cheater
Is capable of
Is located at
(Up until 2020) fully crowdsourced by humans
1.33M commonsense if-then inferences
23 relations (or inference types)
Before, X needs …
Maintain their car
Because X wanted to …
X gets X's car repaired
Is made of
Paper
Fold into origami
Used for
Can be hindered by
The car costs too much
X's car is totaled completely
Happens after
X drives an old car
Happens before
X likes driving now
X spends a fortune
As a result, X reacts
X steals the car
Because X wanted
Make money
X is seen as
Shady
Can be hindered by
X can't find a crowbar
The car would not start
Because X wanted
Risky
Is located at
Gambler
X is seen as
X plays blackjack

Knowledge Models

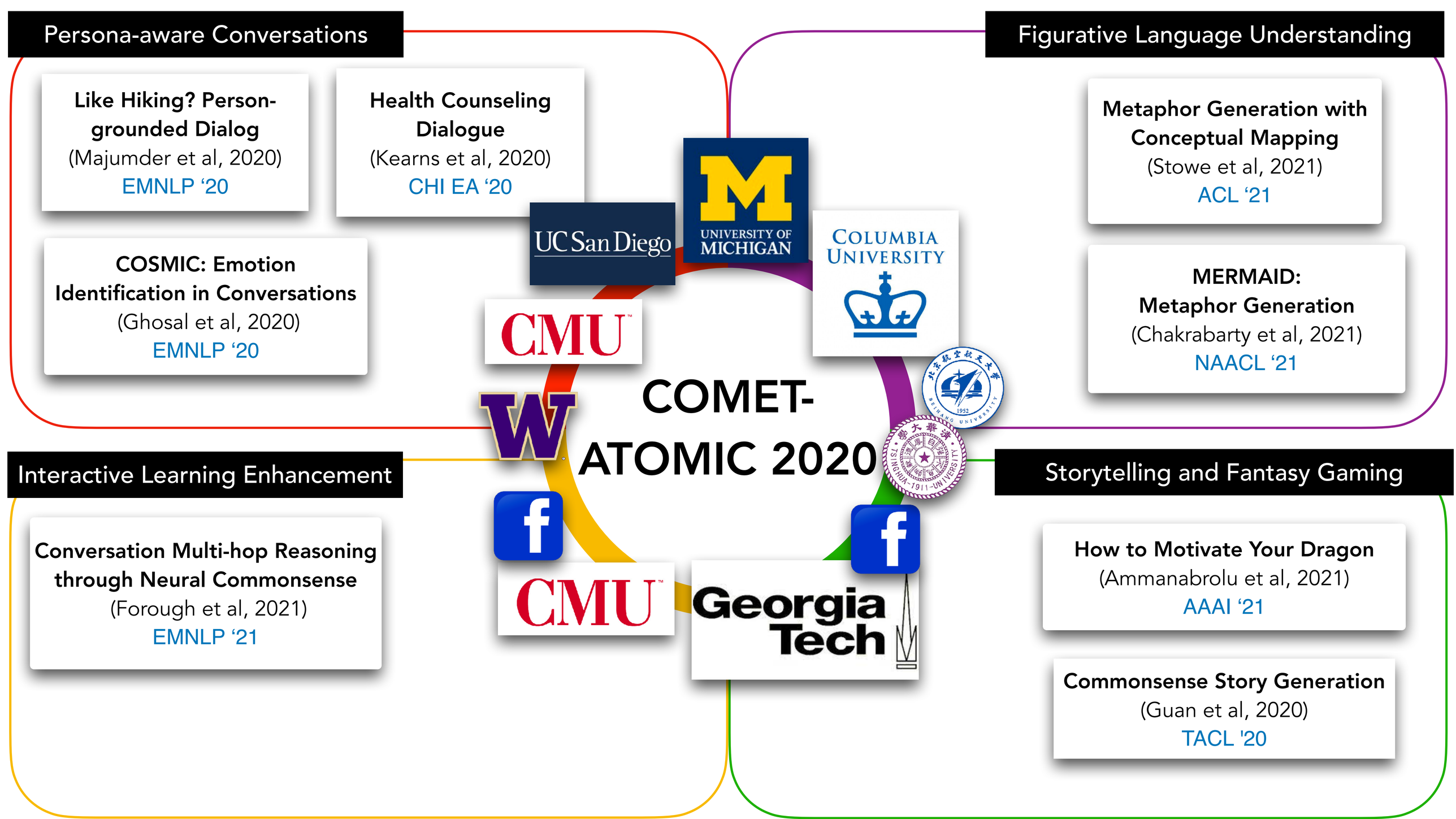
Off-the-shelf Language Models





COMeT (BART): x435 smaller model (~400M parameters), informed by **ATOMIC**$^{20}_{20}$

GPT-3 (Few Shot): 175B parameters!!
pre-trained with a ton of web text (~500B tokens)


Persona-aware Conversations
Like Hiking? Person-grounded Dialog
(Majumder et al, 2020)
EMNLP '20
Health Counseling Dialogue
(Kearns et al, 2020)
CHI EA '20
COSMIC: Emotion Identification in Conversations
(Ghosal et al, 2020)
EMNLP '20
Figurative Language Understanding
Metaphor Generation with Conceptual Mapping
(Stowe et al, 2021)
ACL '21
MERMAID: Metaphor Generation
(Chakrabarty et al, 2021)
NAACL '21
COMET-ATOMIC 2020
Interactive Learning Enhancement
Conversation Multi-hop Reasoning through Neural Commonsense
(Forough et al, 2021)
EMNLP '21
Storytelling and Fantasy Gaming
How to Motivate Your Dragon
(Ammanabrolu et al, 2021)
AAAI '21
Commonsense Story Generation
(Guan et al, 2020)
TACL '20

Symbolic Knowledge Distillation

From Neural Language Models to Causal Commonsense Models

New:
ATOMIC-10x
COMET-distill

Peter West

Chandra Bhagavatula

Jack Hessel

Jena Hwang

Liwei Jiang

Ronan Le Bras

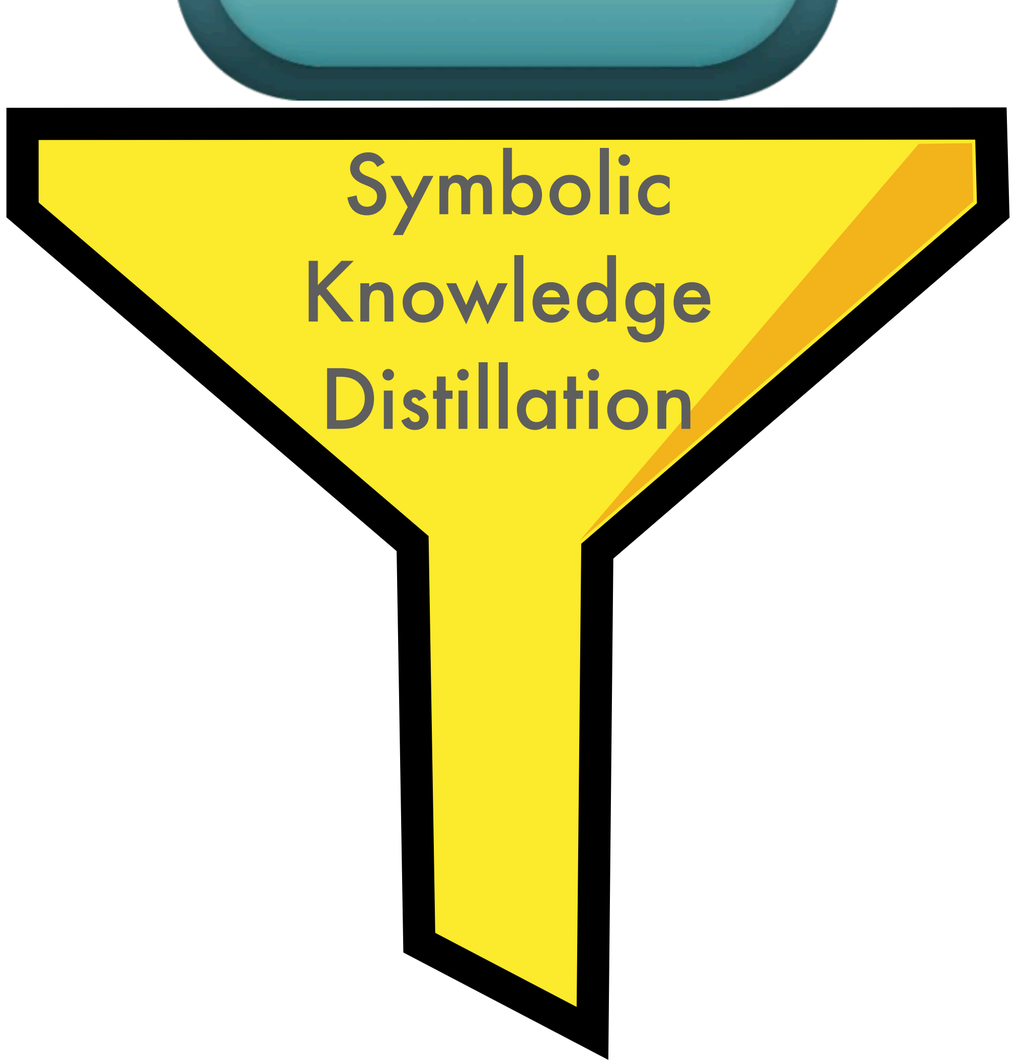
Ximing Lu

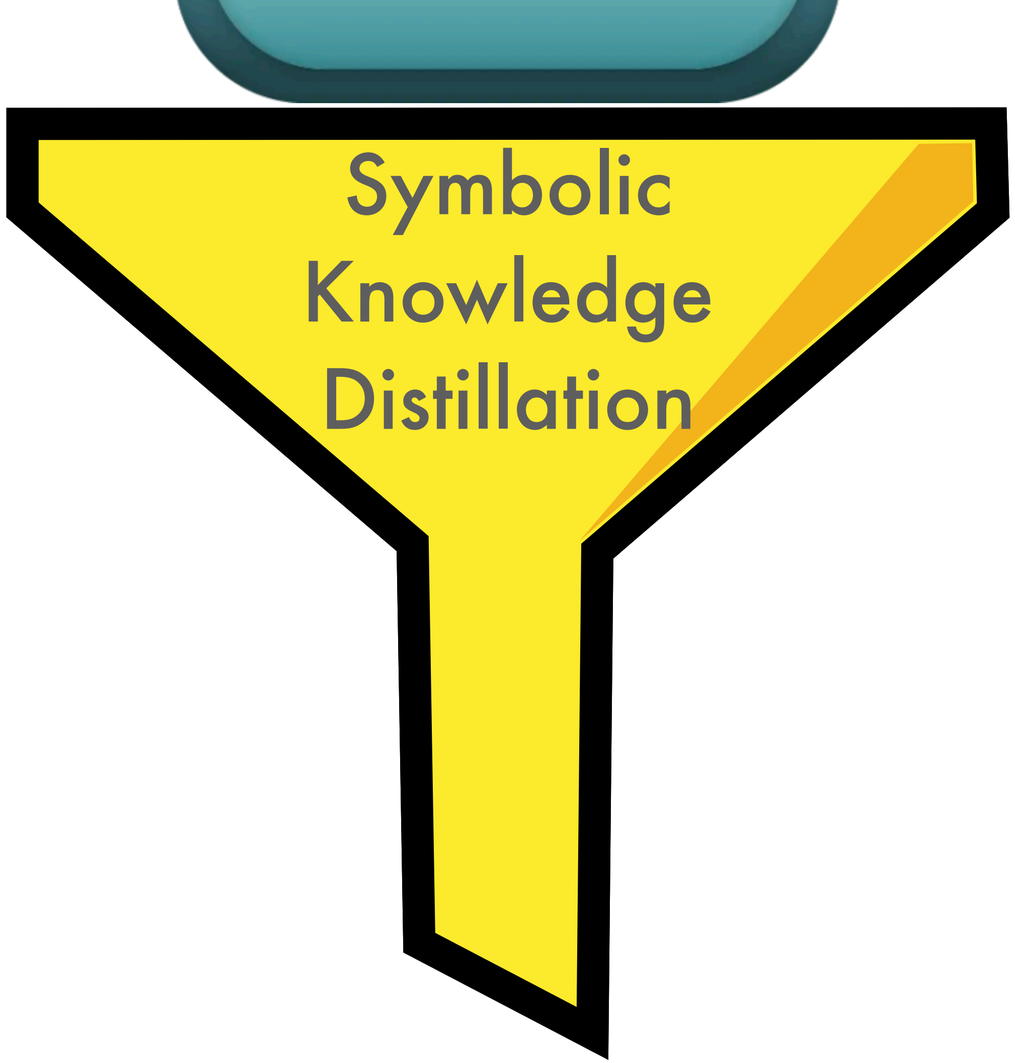
Sean Welleck

Yejin Choi

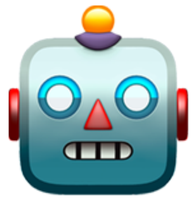
GPT-3

Symbolic
Knowledge
Distillation

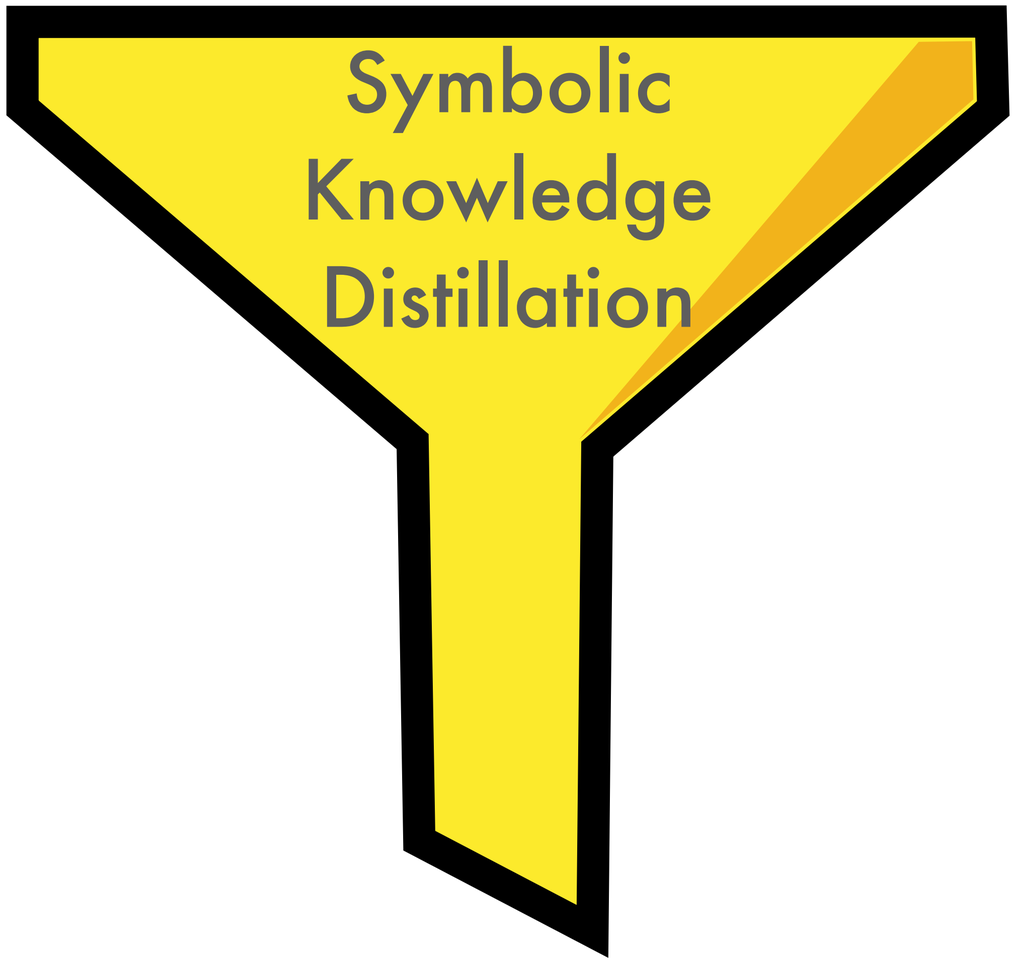

Symbolic
Knowledge
Distillation


Symbolic
Knowledge
Distillation

SMALLER
AND
BETTER



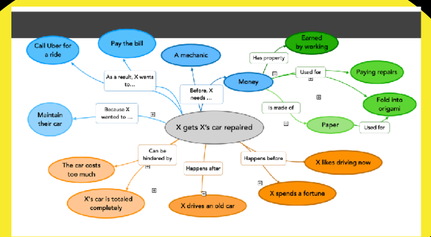
EVEN POSSIBLE
???



Student Model

smaller & better

Symbolic Knowledge Distillation



Knowledge Graph

6.5M high quality examples

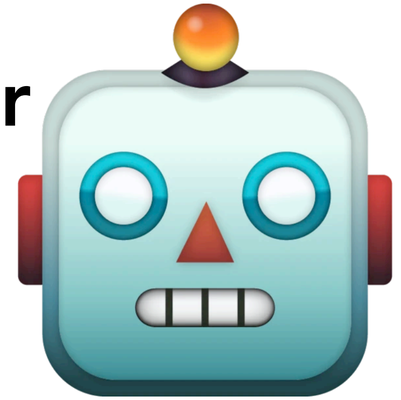
Critic

sort good and bad knowledge

Knowledge Distillation

(Hinton et al. 2015)

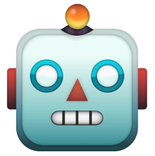
Teacher



$$H(P_t, P_s) = -\sum_{y \in Y} P_t(y) \log P_s(y)$$

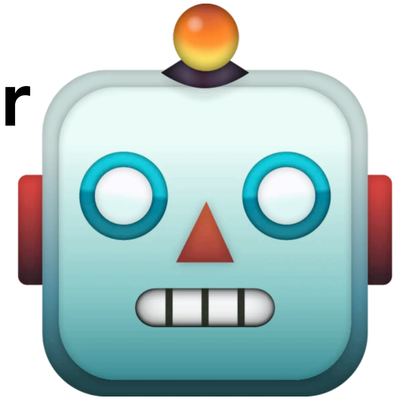
Train student to match teacher probabilities

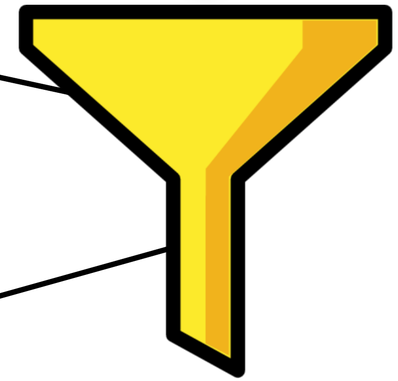
Student



Symbolic
Knowledge Distillation

Teacher

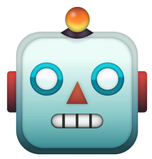




$$H(P_t, P_s) = \sum_{y \in Y} -P_t(y) \log P_s(y)$$

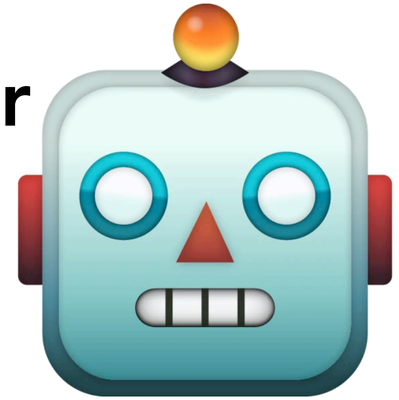
In generation, Y is all strings – intractable!

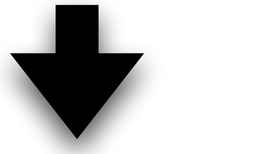
Student

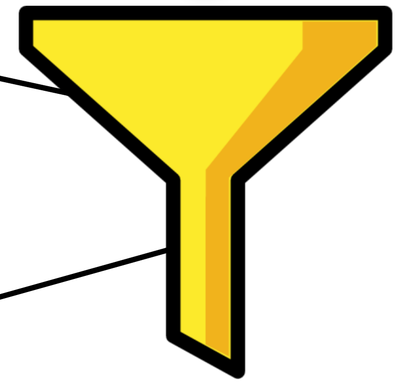


Symbolic Knowledge Distillation

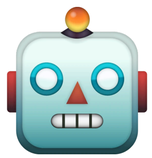
Teacher







Student



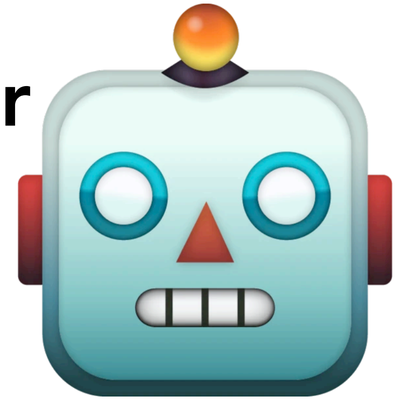
$$H(P_t, P_s) = \mathop{\mathbb{E}}_{y \sim P_t(y)} [-\log P_s(y)]$$

Estimate instead by generating examples!

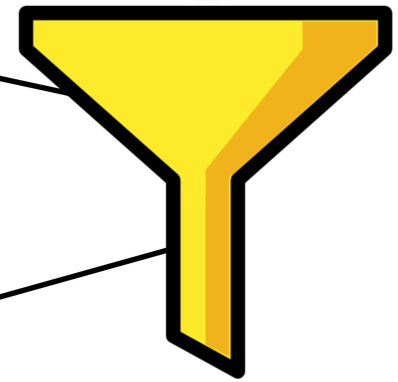
Natural byproduct is a knowledge graph

Symbolic
Knowledge Distillation

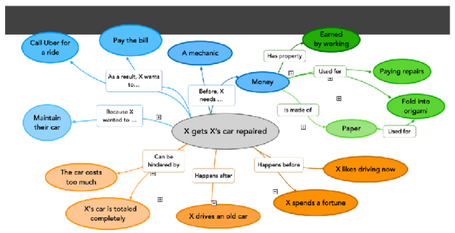
Teacher



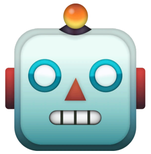
$$H(P_t, P_s) = \mathop{\mathbb{E}}_{y \sim P_t(y)} [-\log P_s(y)]$$



KG



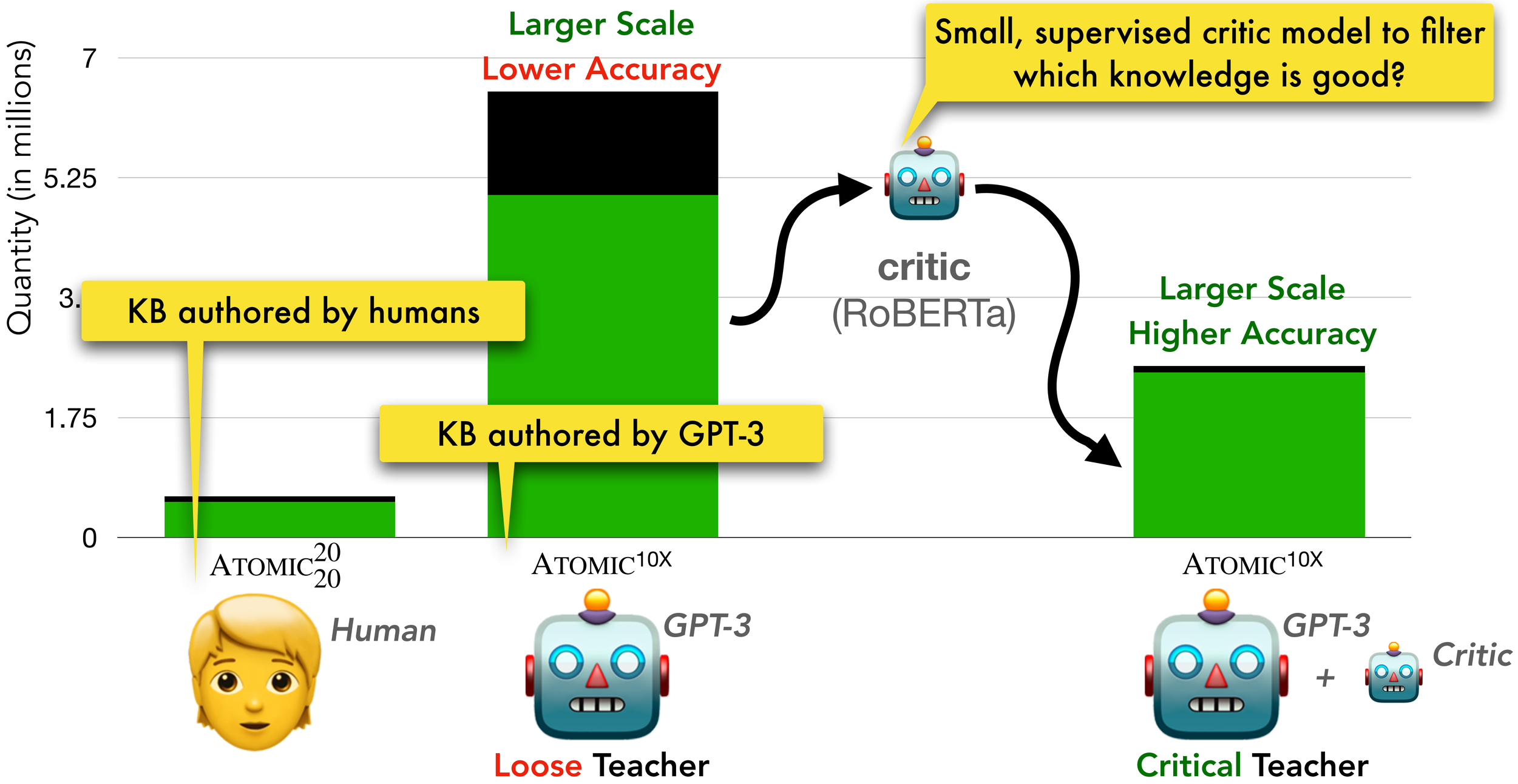
Student

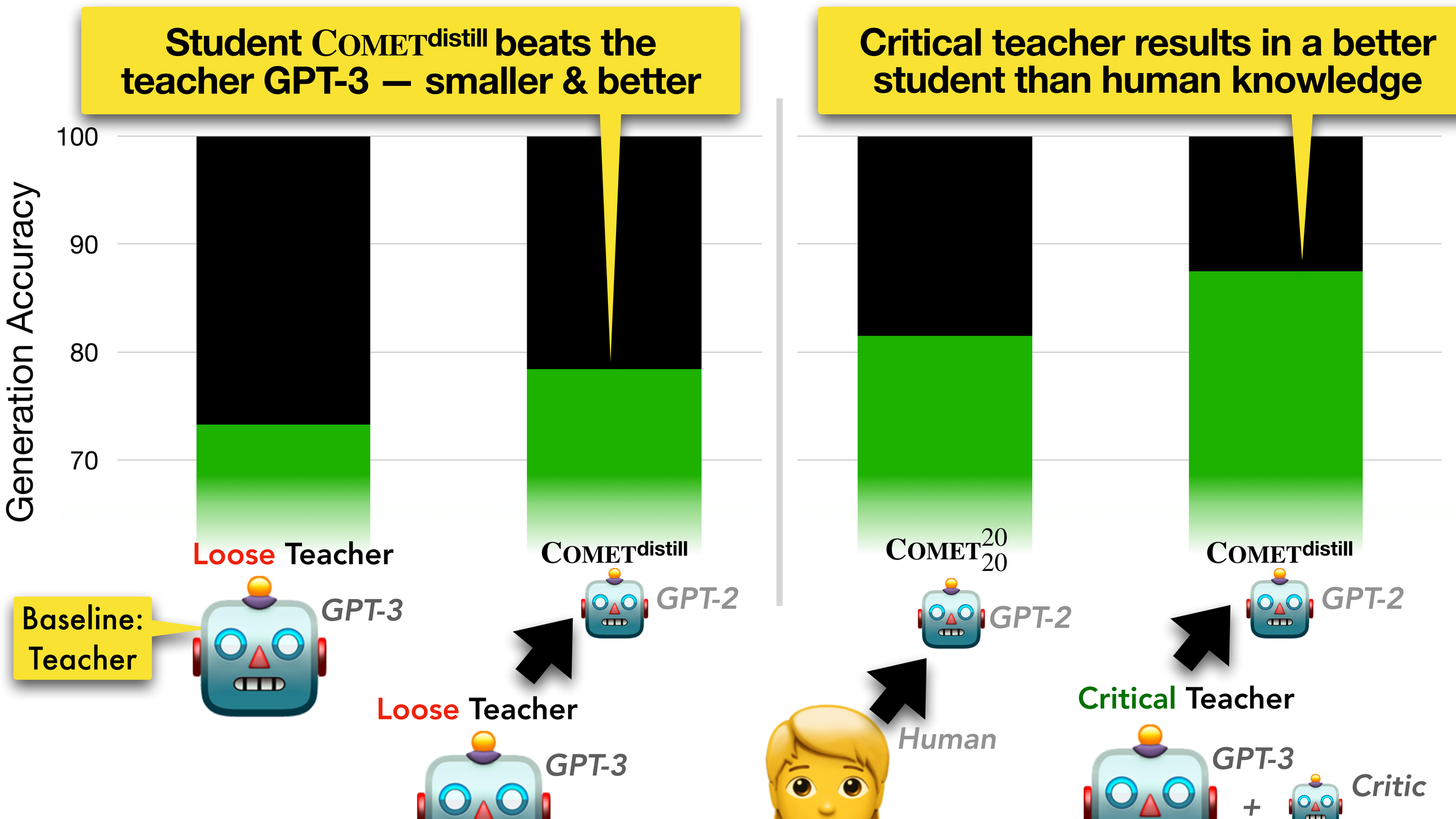


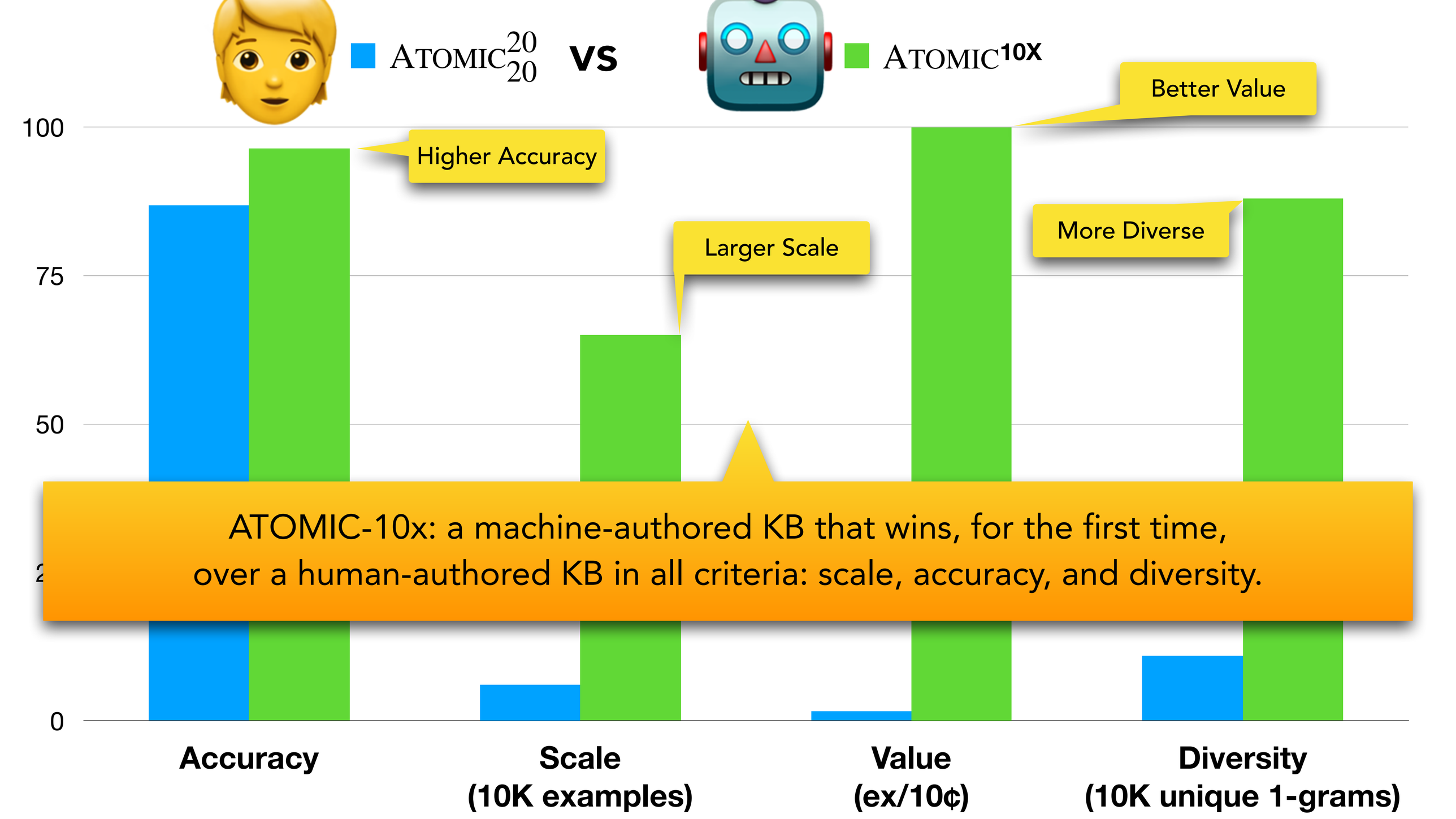
Estimate instead by generating examples!

Natural byproduct is a knowledge graph

Does Symbolic Knowledge Distillation Produce Good knowledge?




Student COMET^distill beats the teacher GPT-3 — smaller & better
Critical teacher results in a better student than human knowledge
Generation Accuracy
100
90
80
70
Loose Teacher
GPT-3
Baseline: Teacher
COMET^distill
GPT-2
Loose Teacher
GPT-3
COMET^20_20
GPT-2
Human
COMET^distill
GPT-2
Critical Teacher
GPT-3
+
Critic


ATOMIC$^{20}_{20}$
vs
ATOMIC^{10X}
Higher Accuracy
Larger Scale
Better Value
More Diverse
100
75
50
0
ATOMIC-10x: a machine-authored KB that wins, for the first time,
over a human-authored KB in all criteria: scale, accuracy, and diversity.
Accuracy
Scale
(10K examples)
Value
(ex/10¢)
Diversity
(10K unique 1-grams)

Thanks! Questions?