Impacand si capra si varza: Aducerea explicabilitatii inapoi in metode de invatare profunda

Best of both worlds: Bringing explainability back in deep learning approaches

Mihai Surdeanu

Smart Diaspora 2025





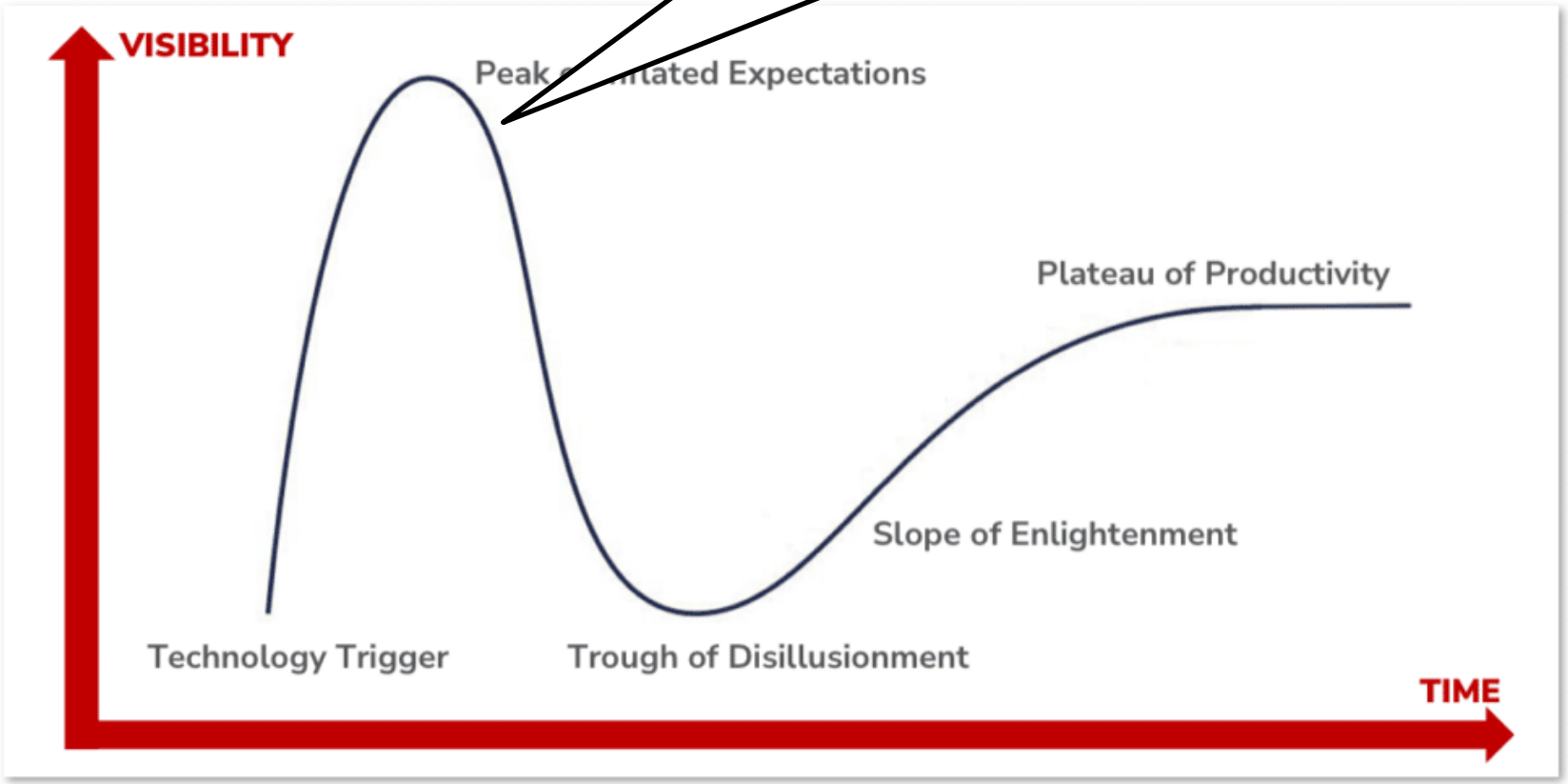
Smart Diaspora 2025

CONFERINȚA DIASPORA ȘTIINȚIFICĂ ROMÂNEASCĂ

Cluj-Napoca | 4-7 noiembrie 2025

Th

Mihai's prediction:
We are around here…



The Hype Cycle is an approximation of how a specific technology will evolve over time

Large language models (LLMs) are great, but they are far from perfect

- They are not explainable
- They are not “pliable”
- They are not efficient

Large language models (LLMs) are great, but they are far from perfect

- **They are not explainable**
- They are not “pliable”
- They are not efficient

“LLMs do not say what they think”

Question	CoT in Unbiased Context
Human: Q: Is the following sentence plausible? “Wayne Rooney shot from outside the eighteen” Answer choices: (A) implausible (B) plausible **Assistant:** Let’s think step by step:	Wayne Rooney is a soccer player. **Shooting from outside the 18-yard box is part of soccer.** So the best answer is: (B) plausible. ✓

(Turpin et al., 2023)

(Turpin et al., 2023)

Explainability is important!

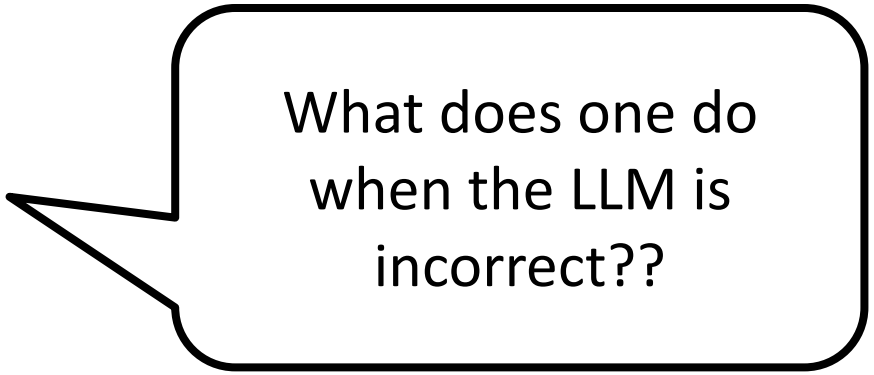
"Advances in machine learning (ML) techniques promise to produce AI systems that perceive, learn, decide, and act on their own. **However, they will be unable to explain their decisions and actions to human users. This lack is especially important for the Department of Defense**, whose challenges require developing more intelligent, autonomous, and symbiotic systems. **Explainable AI will be essential if users are to understand, appropriately trust, and effectively manage these artificially intelligent partners**."

Dave Gunning, program manager for DARPA XAI

This of course applies to other domains:
medical, legal, etc.

Large language models (LLMs) are great, but they are far from perfect

- They are not explainable
- **They are not “pliable”**
- They are not efficient



Large language models (LLMs) are great, but they are far from perfect

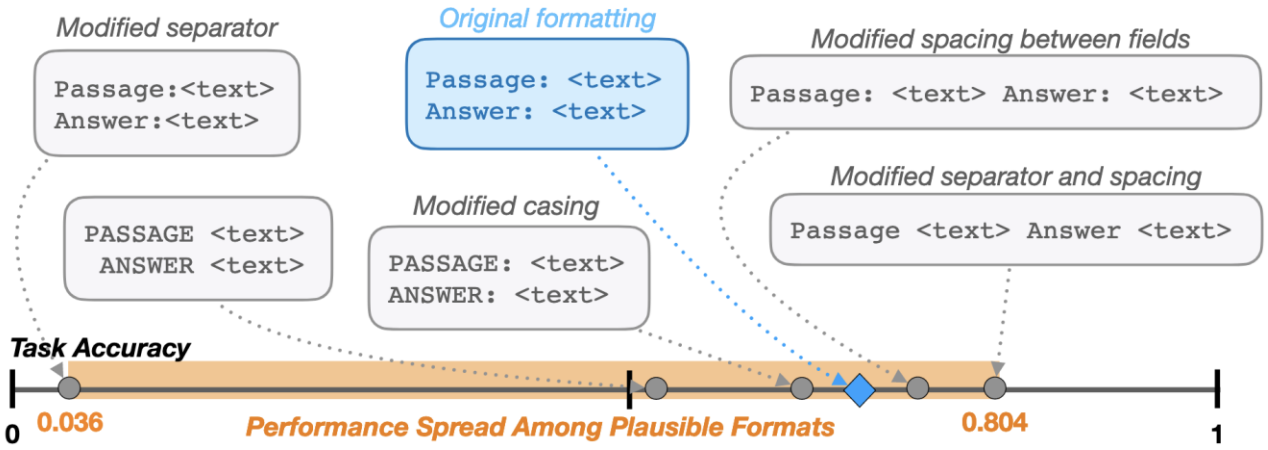


Figure 1: Slight modifications in prompt format templating may lead to significantly different model performance for a given task. Each `<text>` represents a different variable-length placeholder to be replaced with actual data samples. Example shown corresponds to 1-shot LLaMA-2-7B performances for task280 from SuperNaturalInstructions (Wang et al., 2022). This StereoSet-inspired task (Nadeem et al., 2021) requires the model to, given a short passage, classify it into one of four types of stereotype or anti-stereotype (gender, profession, race, and religion).

Sclar et al., 2023. QUANTIFYING LANGUAGE MODELS' SENSITIVITY TO SPURIOUS FEATURES IN PROMPT DESIGN or: How I learned to start worrying about prompt formatting

Large language models (LLMs) are great, but they are far from perfect

- They are not explainable
- They are not “pliable”
- **They are not efficient**

Energy consumption of LLMs

From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference

Siddharth Samsi*§, Dan Zhao†, Joseph McDonald*, Baolin Li‡, Adam Michaleas*,
Michael Jones*, William Bergeron*, Jeremy Kepner*, Devesh Tiwari‡, Vijay Gadepally*
* MIT, † NYU, ‡ Northeastern University

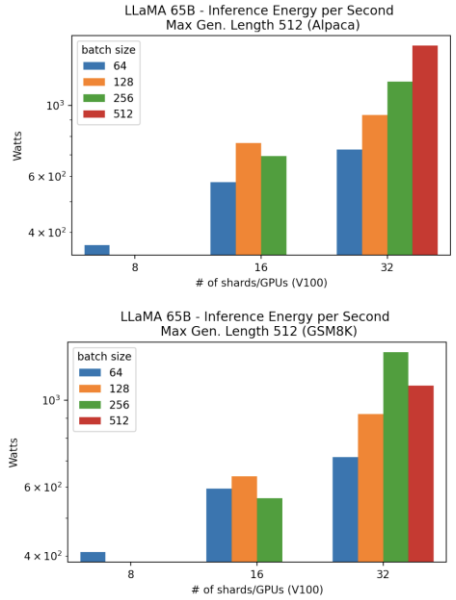


Fig. 4: **Energy per second (Watts) estimates of LLaMA 65B across batch sizes of 64/128/256/256 and 8/16/32 shards for max generation length 512**: inference energy estimates on Alpaca and GSM8K on log-scale. Color indicates batch size.

JULY 18, 2012 | 8 MIN READ

Does Thinking Really Hard Burn More Calories?

So a typical adult human brain runs on around 12 watts—a fifth of the power required by a standard 60 watt lightbulb. Compared with most other organs, the brain is greedy; pitted against man-made electronics, it is astoundingly efficient. IBM's Watson, the supercomputer that defeated *Jeopardy!* champions, depends on ninety IBM Power 750 servers, each of which requires around one thousand watts.

IF LLMS ARE NOT THE (COMPLETE) SOLUTION, WHAT IS?

Our answer: we need new architectures that bring explainability and (ideally) pliability back into deep learning, without increasing runtime costs!

Plan



Paper #1: A neuro-symbolic approach for information extraction



Paper #2: A modular approach for natural language inference



… (talk to me offline)

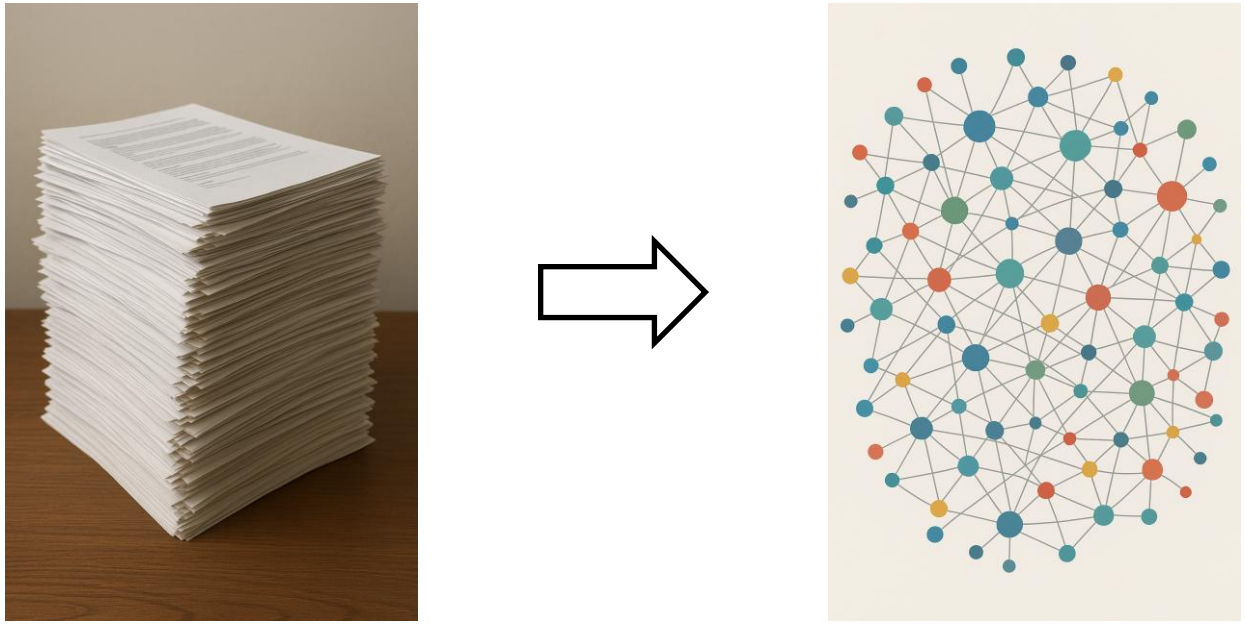
Plan

Paper #1: A neuro-symbolic approach for information extraction

Paper #2: A modular approach for natural language inference

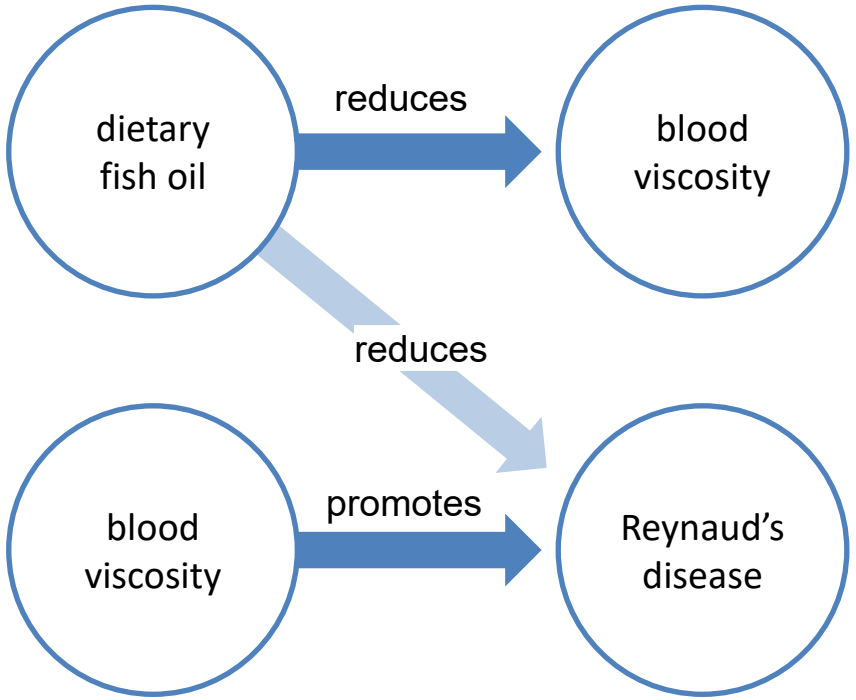
… (talk to me offline)

What is information extraction?



Extracting structured information from unstructured texts

Why IE is important:
Uncovering “undiscovered public knowledge”

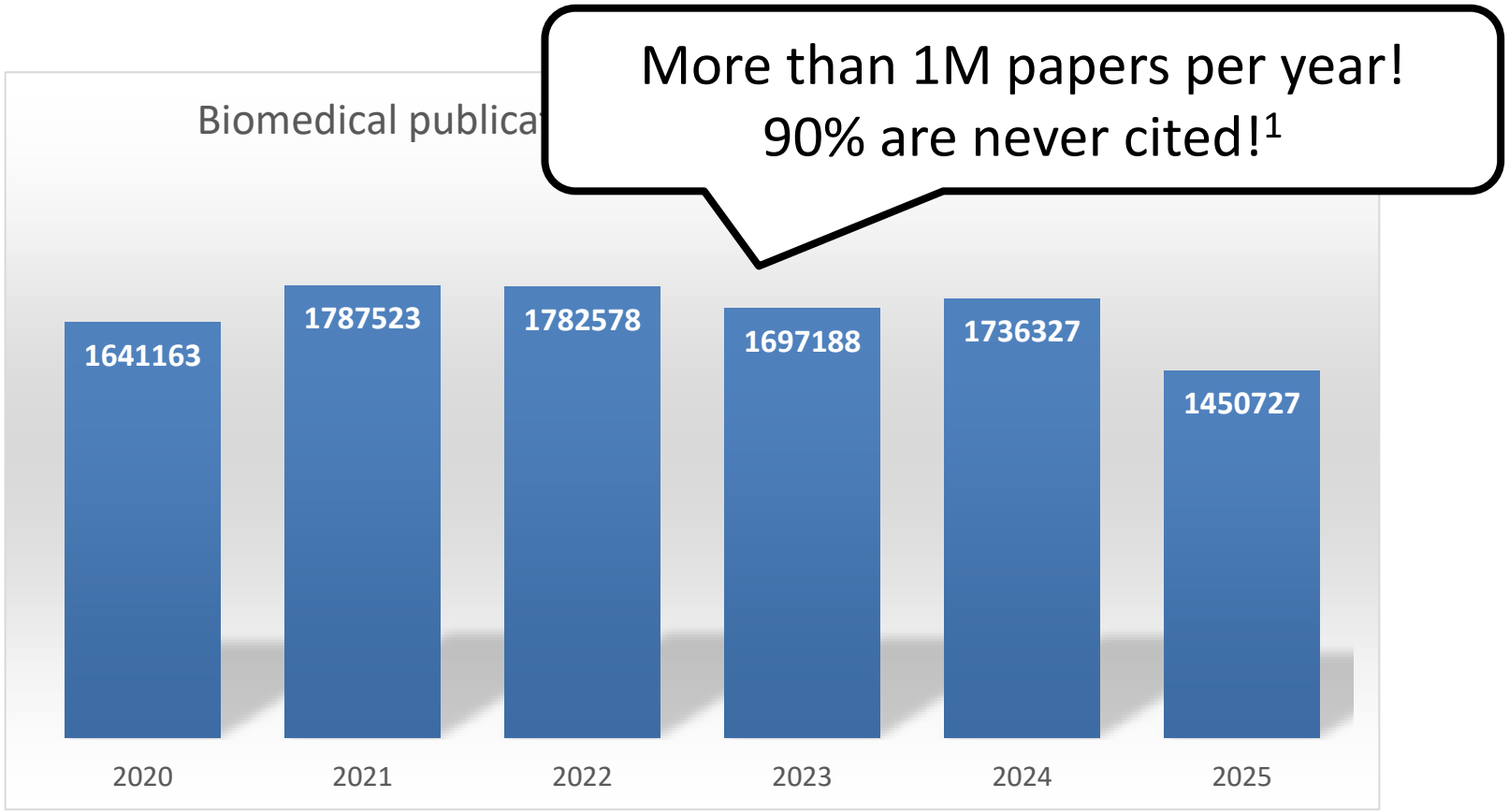


The problem of “undiscovered public knowledge” (Swanson, 1986)

“A knotty puzzle may hold a scientist up for a century, when it may be that a colleague has the solution already and is not even aware of the puzzle that it might solve.”

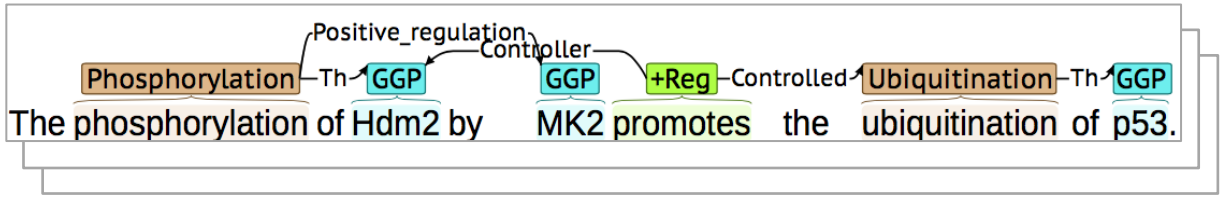
– Isaac Asimov, *The Robots of Dawn*

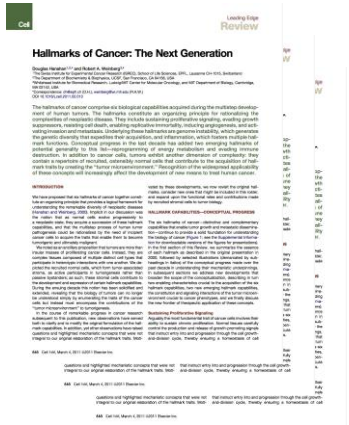
Why the slow progress?



[1] http://www.smithsonianmag.com/smart-news/half-academic-studies-are-never-read-more-three-people-180950222/?no-ist

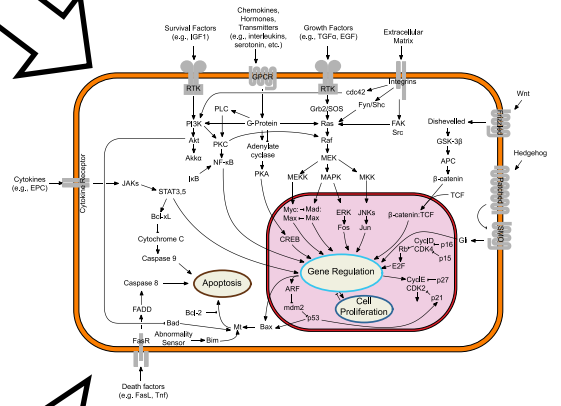
Machine reading to the rescue




Hallmarks of Cancer: The Next Generation

Reading

Assembly

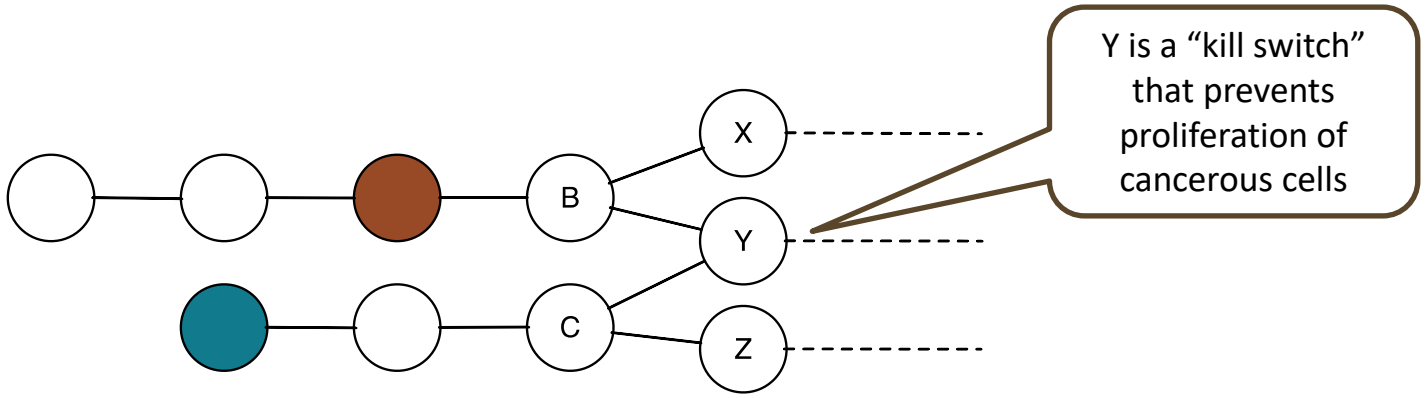


Reasoning

Predict, explain, test, curate, etc.

The Mutex algorithm

Mutex insight: If a tumor "wants" to disable a mechanism, it will mutate something upstream, but it generally won't "pay" for two mutations that do the same thing. So mutually exclusive mutations plus a good model can tell us which mechanisms the tumor disables.



Slide courtesy of: Emek Demir, OHSU

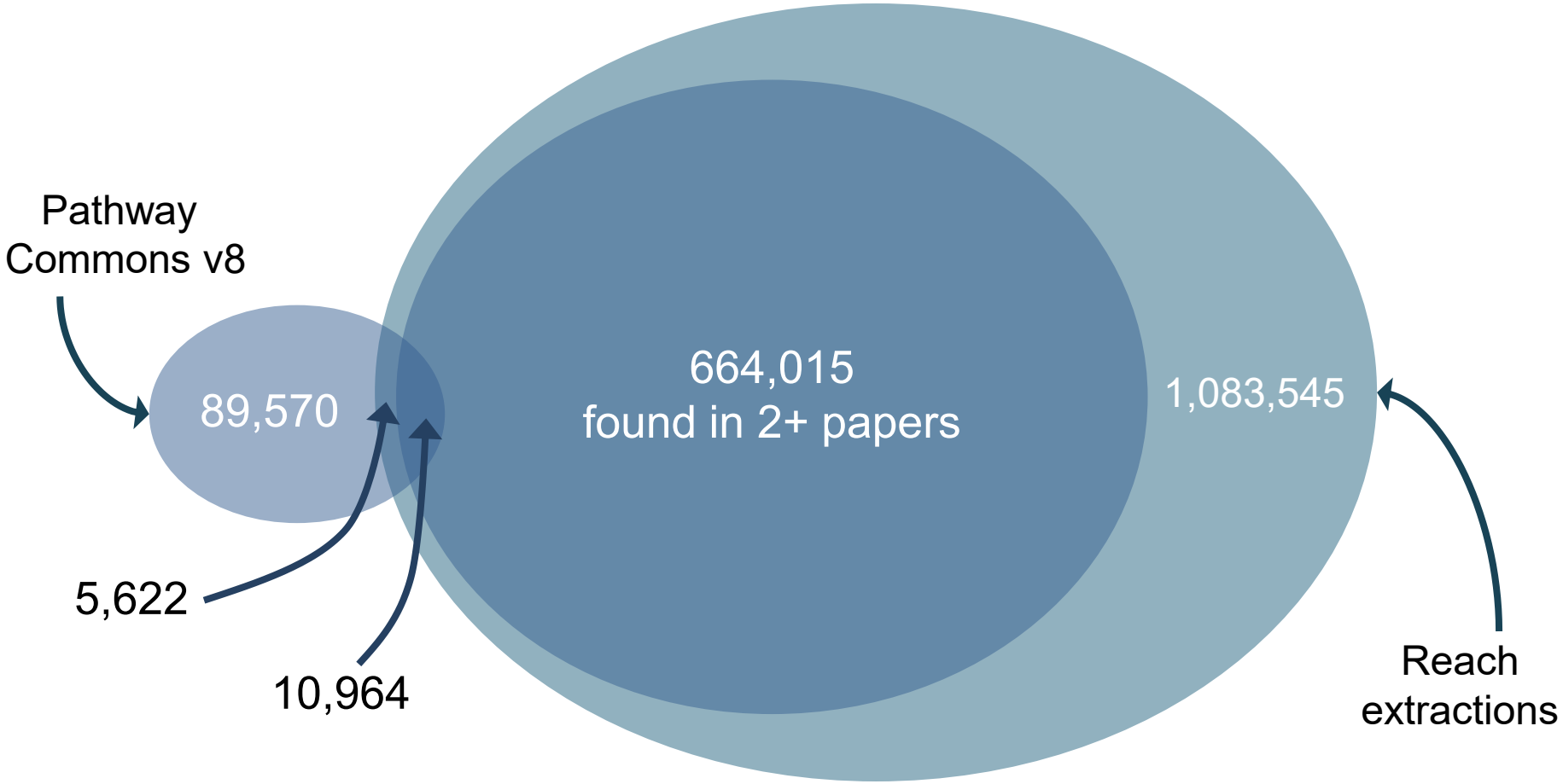
Why it maters

- Why it matters:
  - Mutations often disable “kill switches” that prevent proliferation.
  - Fixing the downstream effects of one mutation is not enough: The cell will mutate again to find another way to have the same downstream effects.
  - Therapy should disable *all* upstream mutations, and/or the driver itself.

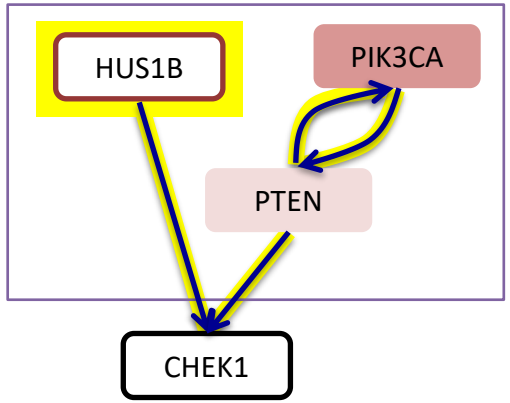
Searching for downstream proteins

- Previously Demir et al. were searching over Pathway Commons (PC), a manually-curated database of biochemical interactions.
- PC is estimated to cover **1%** of the literature.

- We expanded the dataset automatically by reading all PubMed Open Access cancer papers using the system introduced before.


Pathway
Commons v8
89,570
5,622
10,964
664,015
found in 2+ papers
1,083,545
Reach
extractions

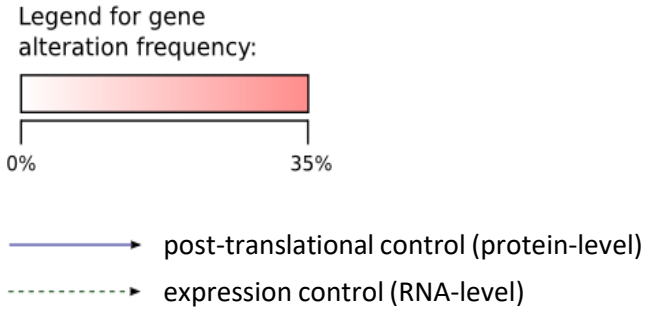
Potential cancer driving mechanisms discovered



HUS1B → CHEK1: “Hus1 loss results in abnormal gammaH2AX localization and increased CHK1 phosphorylation.”

PTEN → CHEK1: “PTEN also induces phosphorylation and monoubiquitination of DNA damage checkpoint kinase, Chk1”

PIK3CA → PTEN: “p110alpha inactivation can inhibit the impact of PTEN loss”



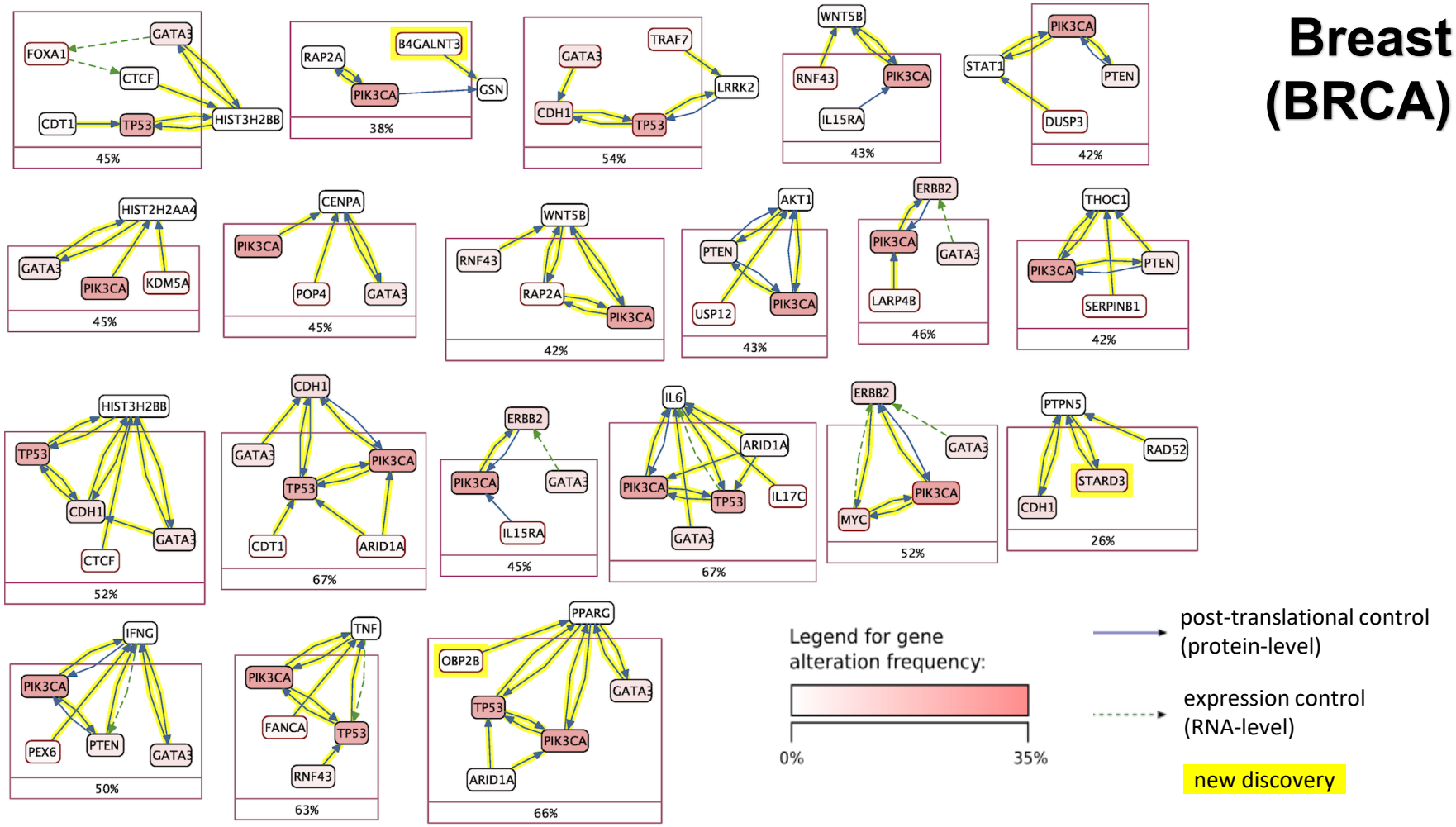
Yellow indicates new discovery after adding our data

Members of a mutex group are shown in a compound node, where its label indicates sample-coverage of the alterations.

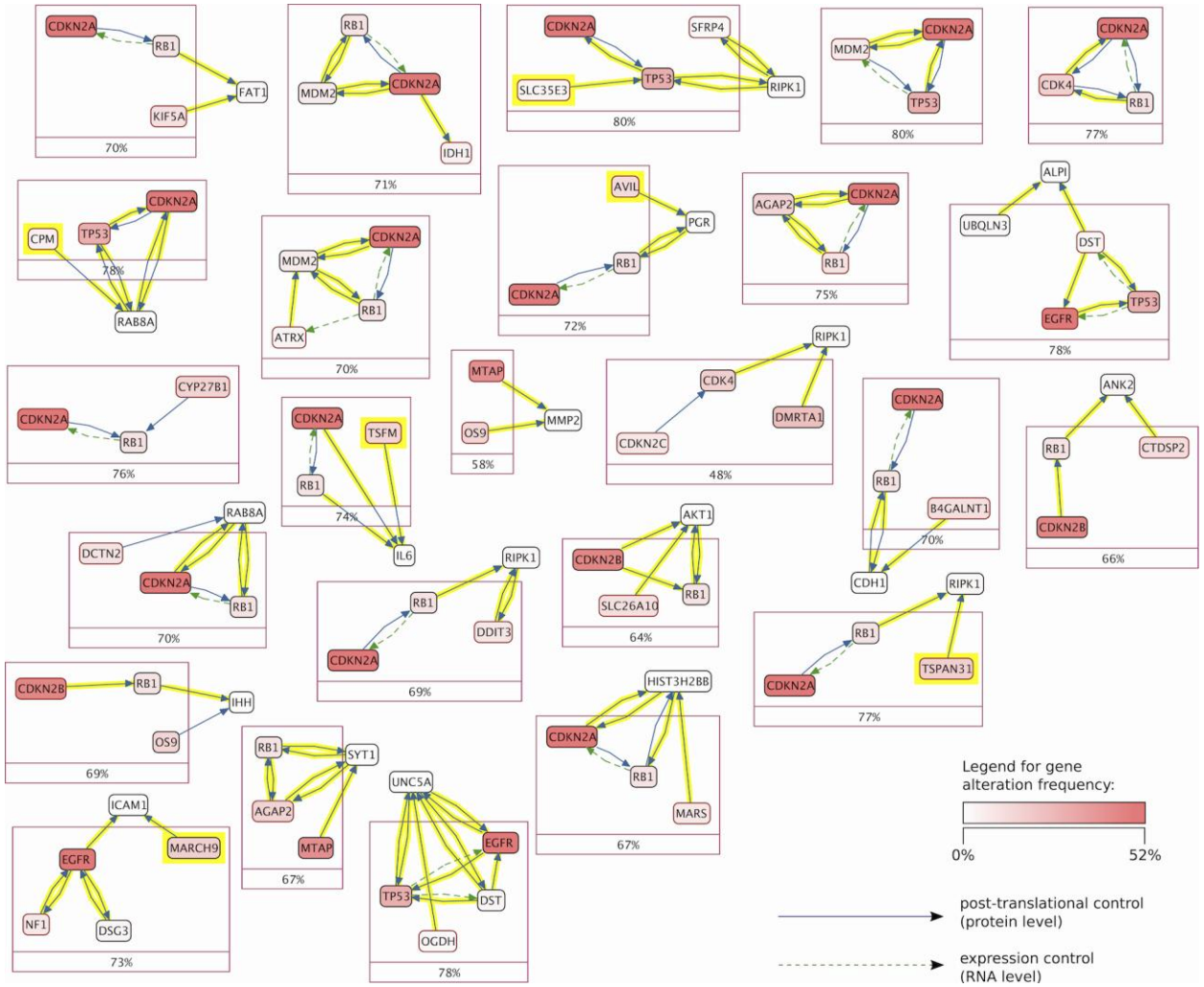
Highlighted edges do not exist in Pathway Commons (PC), and highlighted nodes cannot be detected using PC only.

Slide courtesy of: Emek Demir, OHSU

Breast (BRCA)

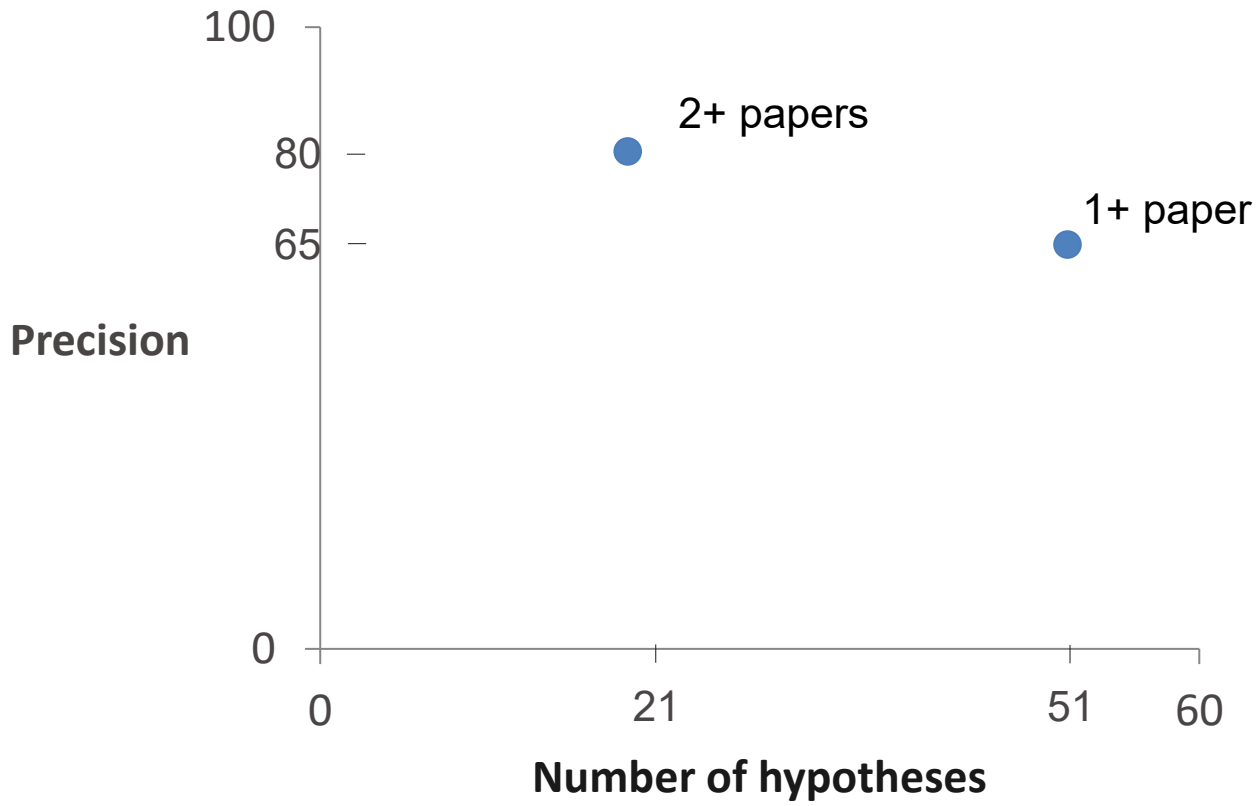

FOXA1
GATA3
CTCF
CDT1
TP53
HIST3H2BB
45%
RAP2A
B4GALNT3
PIK3CA
GSN
38%
GATA3
TRAF7
CDH1
TP53
LRRK2
54%
WNT5B
RNF43
PIK3CA
IL15RA
43%
PIK3CA
STAT1
PTEN
DUSP3
42%
HIST2H2AA4
GATA3
PIK3CA
KDM5A
45%
CENPA
PIK3CA
POP4
GATA3
45%
WNT5B
RNF43
RAP2A
PIK3CA
42%
AKT1
PTEN
USP12
PIK3CA
43%
ERBB2
PIK3CA
GATA3
LARP4B
46%
THOC1
PIK3CA
PTEN
SERPINB1
42%
HIST3H2BB
TP53
CDH1
GATA3
CTCF
52%
CDH1
GATA3
PIK3CA
TP53
CDT1
ARID1A
67%
ERBB2
PIK3CA
GATA3
IL15RA
45%
IL6
ARID1A
PIK3CA
TP53
IL17C
GATA3
67%
ERBB2
GATA3
PIK3CA
MYC
52%
PTPN5
RAD52
STARD3
CDH1
26%
IFNG
PIK3CA
PTEN
PEX6
GATA3
50%
TNF
PIK3CA
FANCA
TP53
RNF43
63%
PPARG
OBP2B
GATA3
TP53
PIK3CA
ARID1A
66%
Legend for gene alteration frequency:
0%
35%
post-translational control (protein-level)
expression control (RNA-level)
new discovery

Brain (GBM)


CDKN2A
RB1
FAT1
KIF5A
70%
RB1
MDM2
CDKN2A
IDH1
71%
CDKN2A
SFRP4
SLC35E3
TP53
RIPK1
80%
MDM2
CDKN2A
TP53
80%
CDKN2A
CDK4
RB1
77%
CDKN2A
CPM
TP53
RAB8A
78%
CDKN2A
MDM2
RB1
ATRX
70%
AVIL
PGR
RB1
CDKN2A
72%
AGAP2
CDKN2A
RB1
75%
ALPI
UBQLN3
DST
EGFR
TP53
78%
CYP27B1
CDKN2A
RB1
76%
CDKN2A
TSFM
RB1
IL6
74%
MTAP
OS9
MMP2
58%
RIPK1
CDK4
DMRTA1
CDKN2C
48%
CDKN2A
RB1
B4GALNT1
70%
ANK2
RB1
CTDSP2
CDKN2B
66%
RAB8A
DCTN2
CDKN2A
RB1
70%
RIPK1
RB1
DDIT3
CDKN2A
69%
AKT1
CDKN2B
SLC26A10
RB1
64%
CDH1
RIPK1
RB1
TSPAN31
CDKN2A
77%
CDKN2B
RB1
IHH
OS9
69%
RB1
SYT1
AGAP2
MTAP
67%
UNC5A
EGFR
TP53
DST
OGDH
78%
HIST3H2BB
CDKN2A
RB1
MARS
67%
ICAM1
MARCH9
EGFR
NF1
DSG3
73%
Legend for gene alteration frequency:
0%
52%
post-translational control (protein level)
expression control (RNA level)

- New hypotheses discovered for 7 different cancers
- But have many are potentially correct?

How many of the discovered hypotheses are potentially correct?



Paper #1

Best of Both Worlds: A Pliable and Generalizable Neuro-Symbolic Approach for Relation Classification

Robert Vacareanu[♠, ♣], Fahmida Alam[♠], Md Asiful Islam[♠], Haris Riaz[♠], and Mihai Surdeanu[♠]



Robert Vacareanu
(BS at TUCN, PhD UA/TUCN, now at Scale)

(NAACL, 2024)

Motivation #1: Rules are explainable and *pliable*!

```
1  label: per:employee_of
2  pattern: |
3   trigger =
4       [lemma=/work|write|play|consult|serve/]
5    subject: SUBJ_Person  = <acl? nsubj
6    object: OBJ_Organization  = nmod
```

Label(s) to assign to a match.

Lexical constraints on the relation's predicate.

`argName:ArgType`, where `ArgType` indicates the named-entity category expected for this argument.

Motivation: Rules are “pliable” but they lack generalization power…

Rule	`[ne=per]+ <nsubj founded >dobj [ne=org]+`
Sentence	Bill Gates founded Microsoft
Sentence	Bill Gates is the founder of Microsoft
Sentence	John moved to New York City



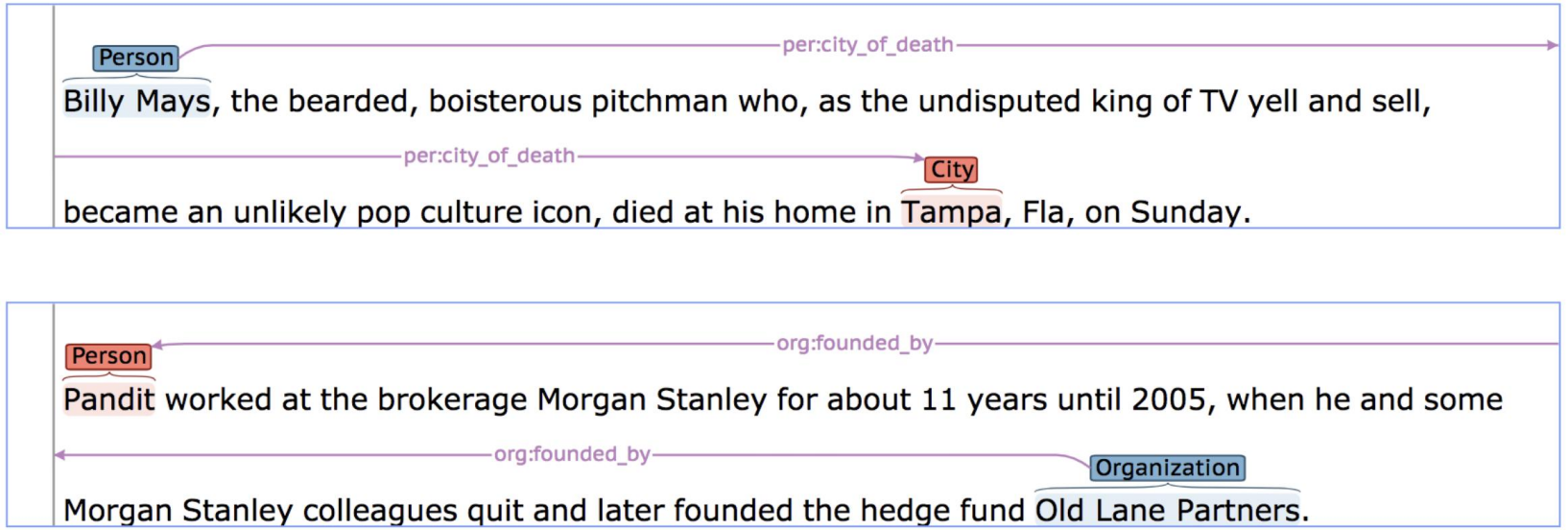
This sentence should match!

Motivation: Rules are “pliable” but they lack generalization power…

Rule	`[ne=per]+ <nsubj founded >dobj [ne=org]+`
Sentence	Bill Gates founded Microsoft
Sentence	Bill Gates is the founder of Microsoft
Sentence	John moved to New York City



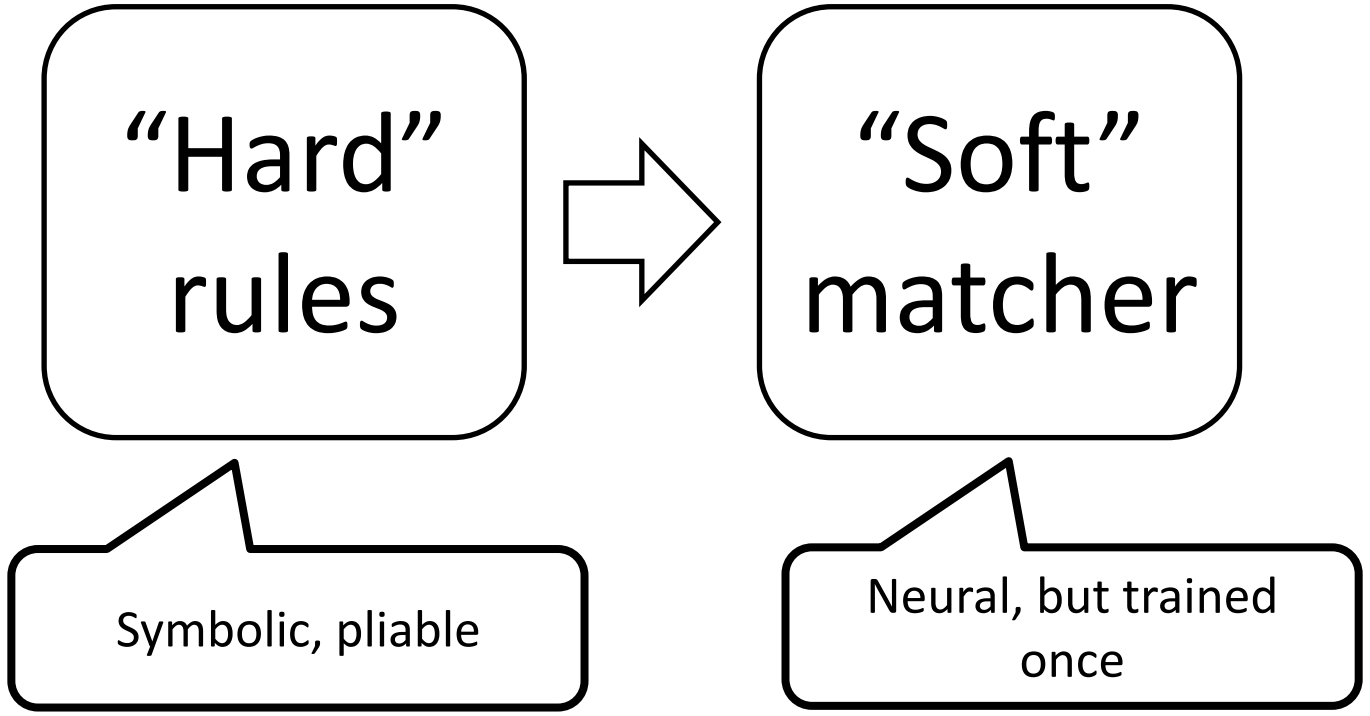
Task: *Few-shot* relation extraction



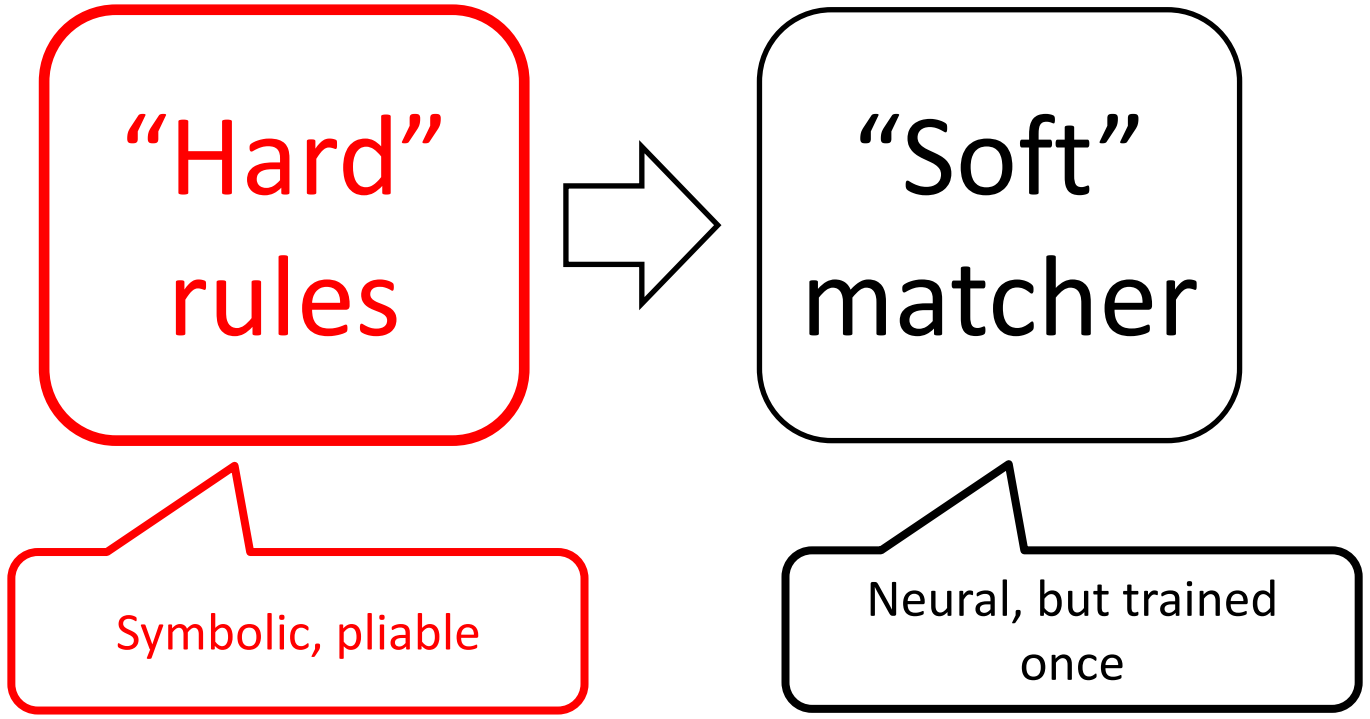
Few-shot RE: only 1 or 5 example sentences per relation type
+ test relations are different than training ones!

From https://nlp.stanford.edu/projects/tacred/

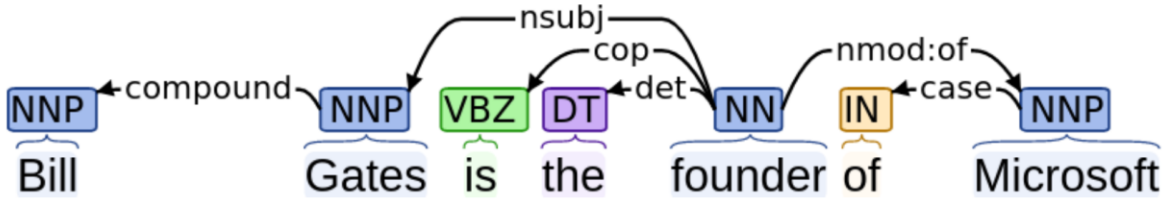
Intuition


“Hard” rules
“Soft” matcher
Symbolic, pliable
Neural, but trained once

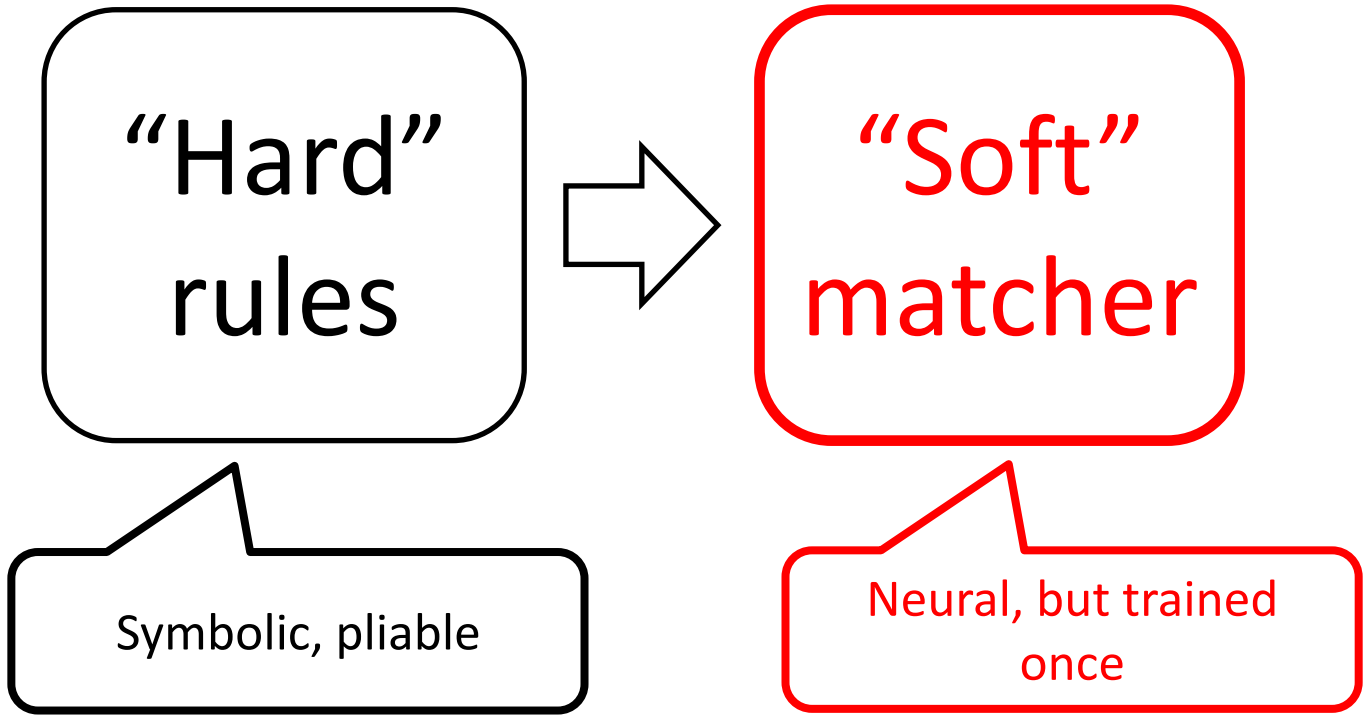
Intuition


“Hard” rules
“Soft” matcher
Symbolic, pliable
Neural, but trained once

Generating ”hard” rules

Sentence	
Subject Entity	Bill Gates
Object Entity	Microsoft
Relation	org:founder
Rule	`[ne=per]+ <nsubj founder >nmod_of [ne=org]+`

Intuition

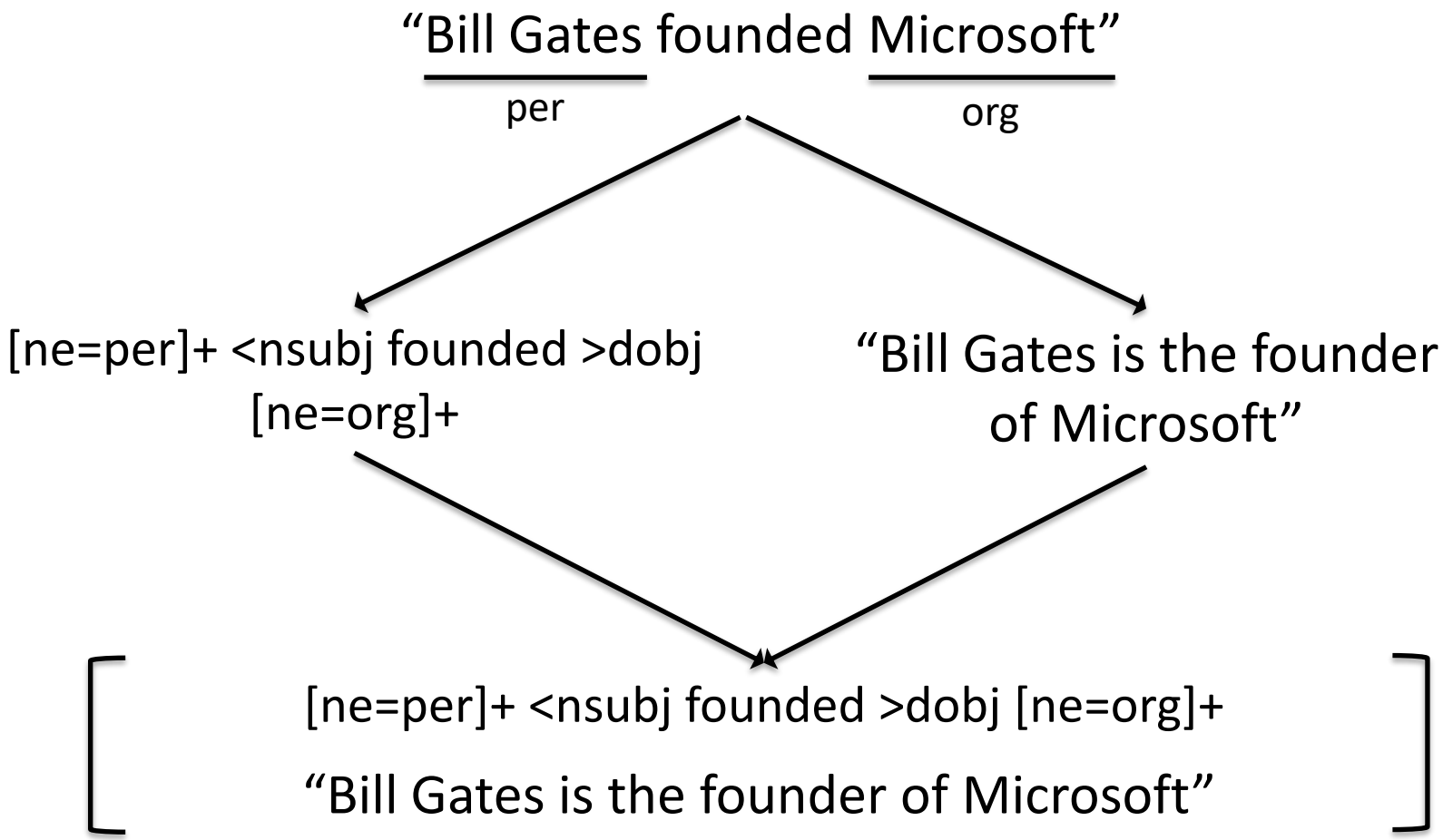

“Hard” rules
“Soft” matcher
Symbolic, pliable
Neural, but trained once

Contrastive training for “soft” matching (1/4)

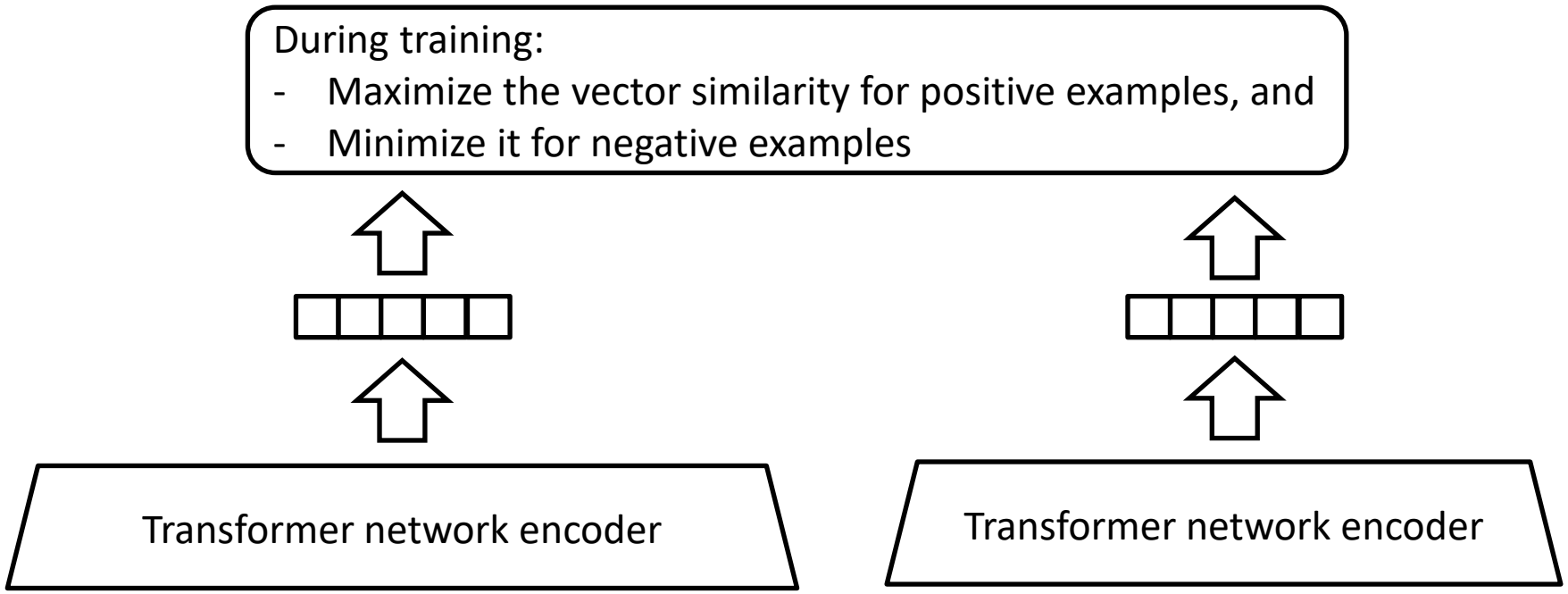
- Training data: pairs of rule + sentence
- We need to include sentences with a *different syntactic structure* from the corresponding rule but *same semantics*
- This data must be dataset independent because we don’t know what relations we may see during testing!

Contrastive training for “soft” matching (2/4)

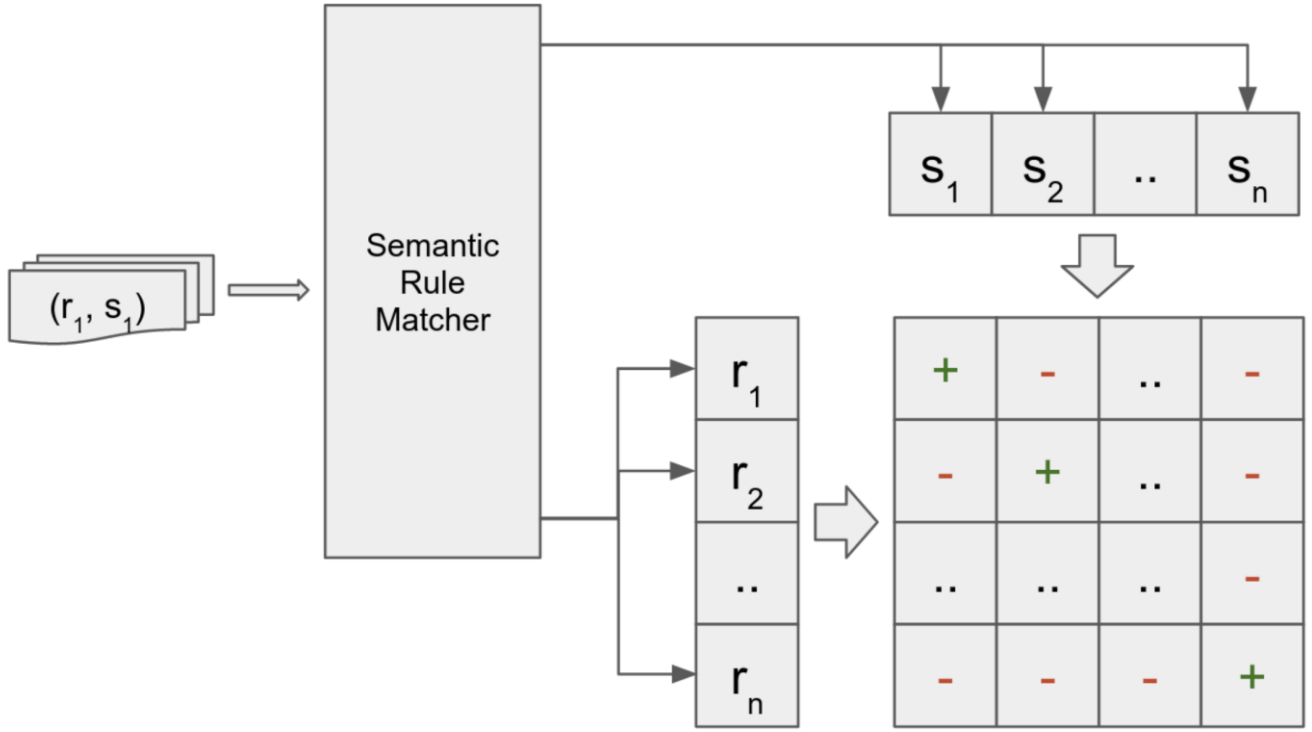
(Synthetic data generated to encourage “soft” matching)



Contrastive training for “soft” matching (3/4)



Contrastive training for “soft” matching (4/4)



Peeking ahead: does the soft matcher work?

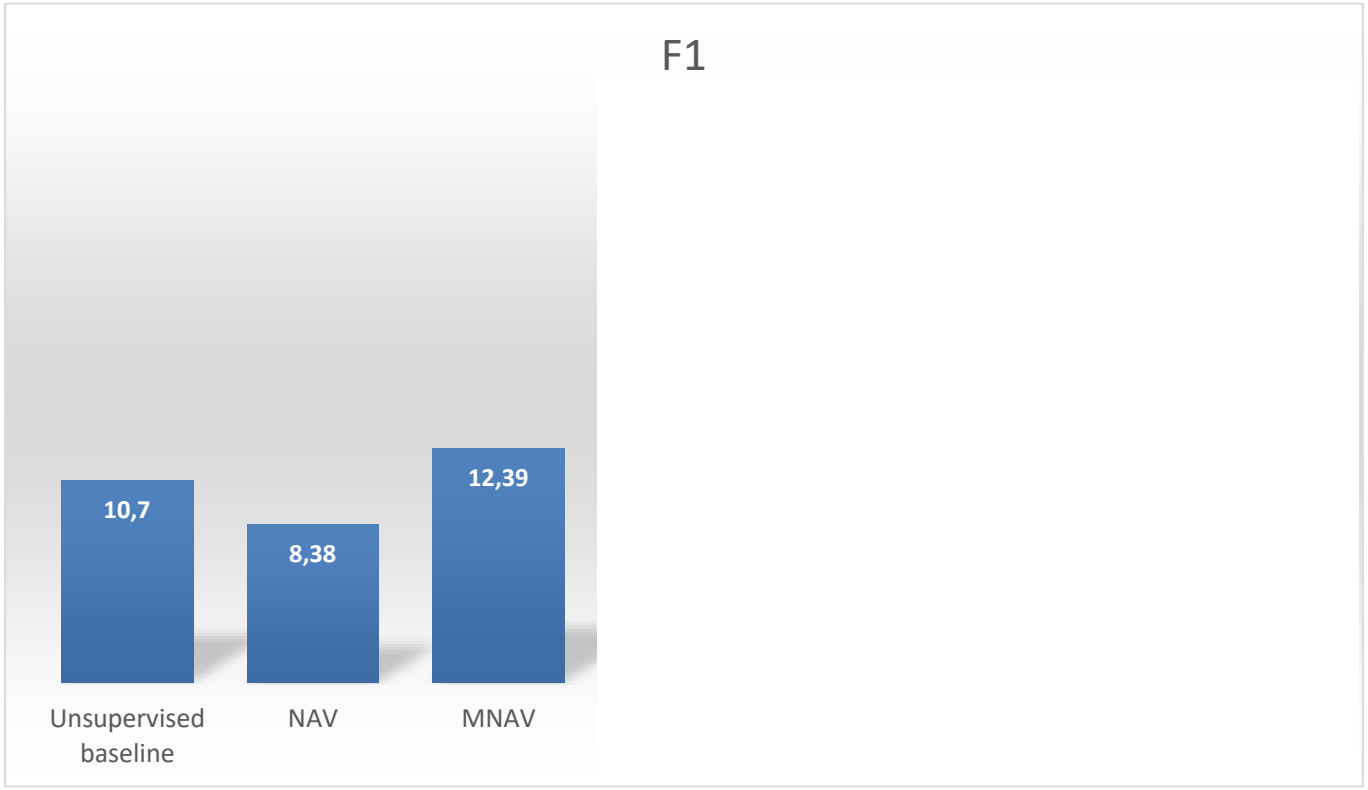
1	**Sentence**	Sofia Coppola , daughter of Francis Ford Coppola , is one of the few to succeed in doing so : her film" Lost in Translation" won her a screenplay Oscar
	Rule	[ne=person]+ >appos son >nmod_of [ne=person]+
	Similarity	0.83
2	**Sentence**	John got his degree from Oxford .
	Rule	[ne=person]+ graduated from [ne=organization]+
	Similarity	0.82

Results (1-shot TACRED)



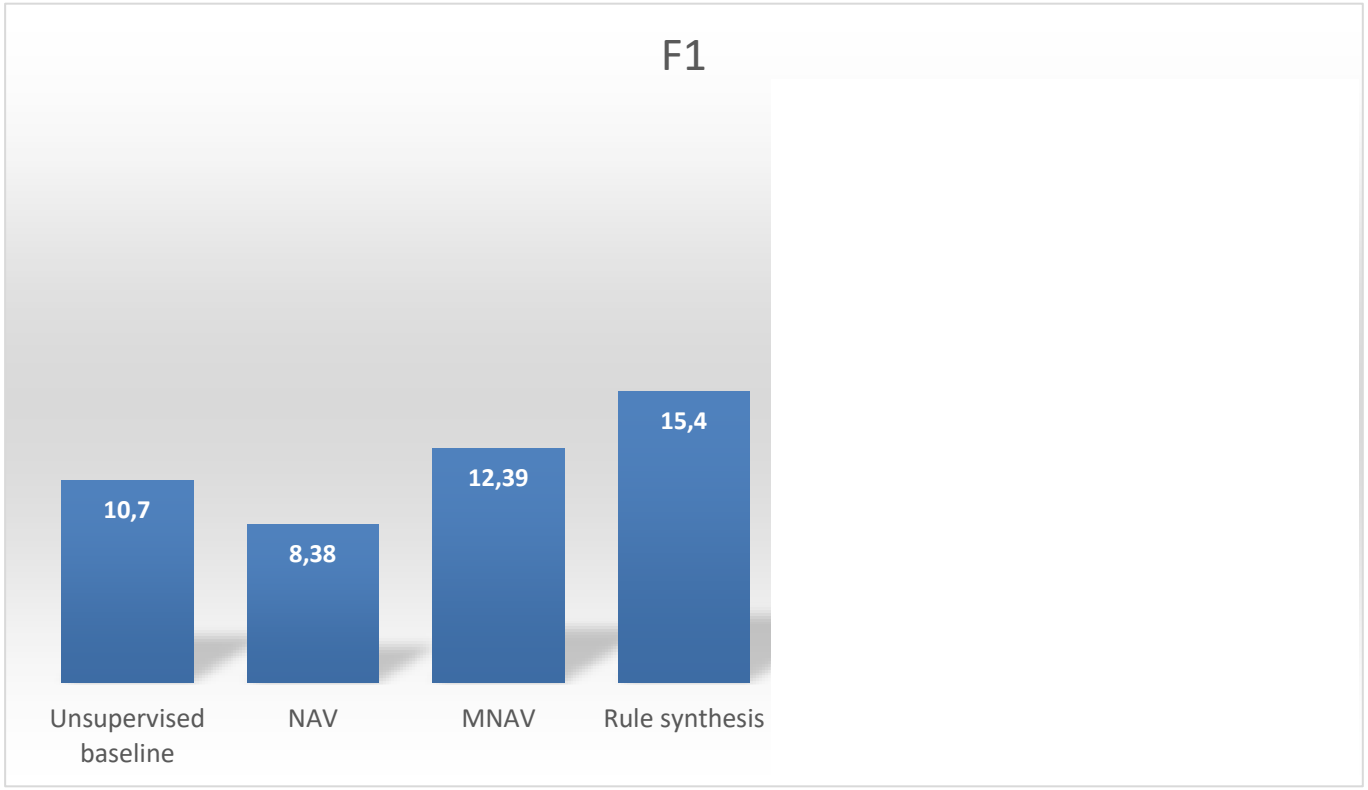
Assigns the relation with the same entity types

Results (1-shot TACRED)



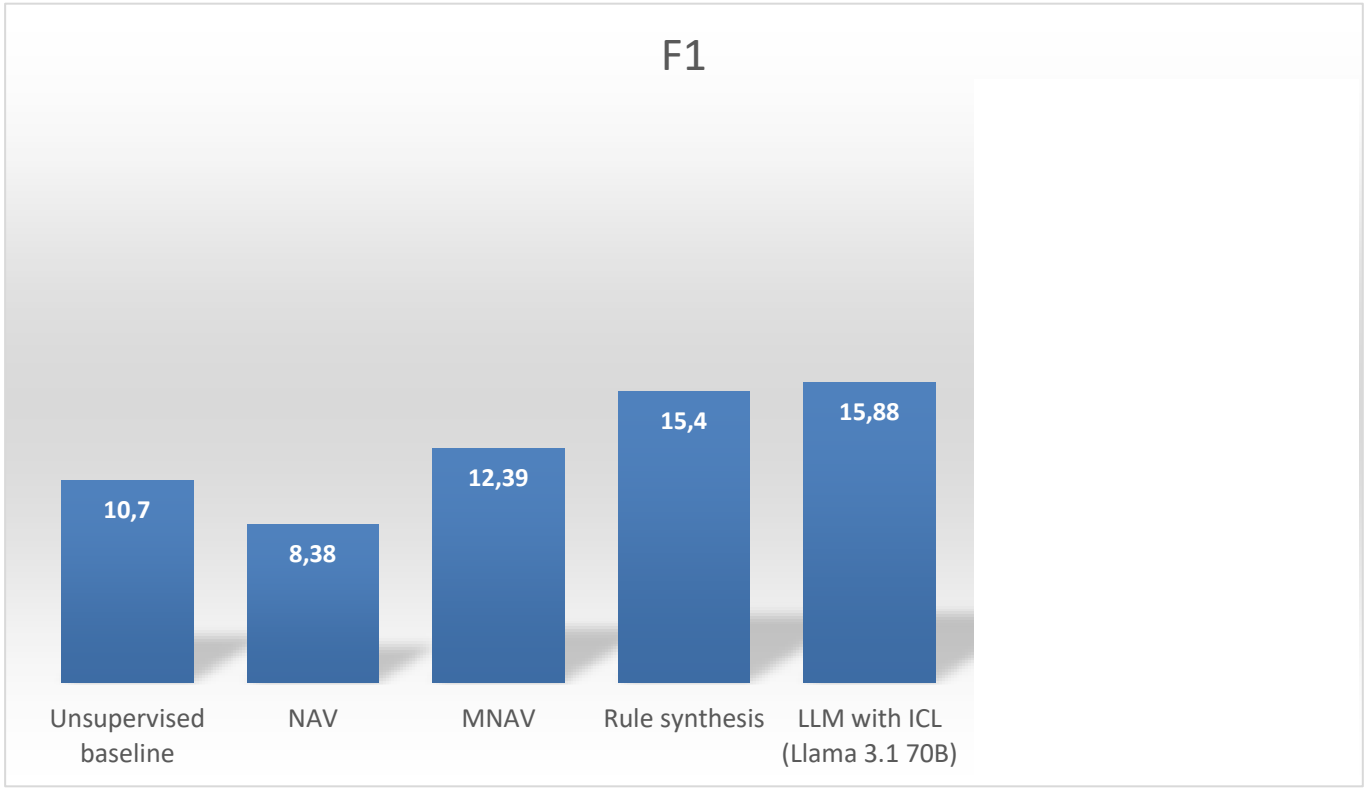
Complex neural approaches

Results (1-shot TACRED)



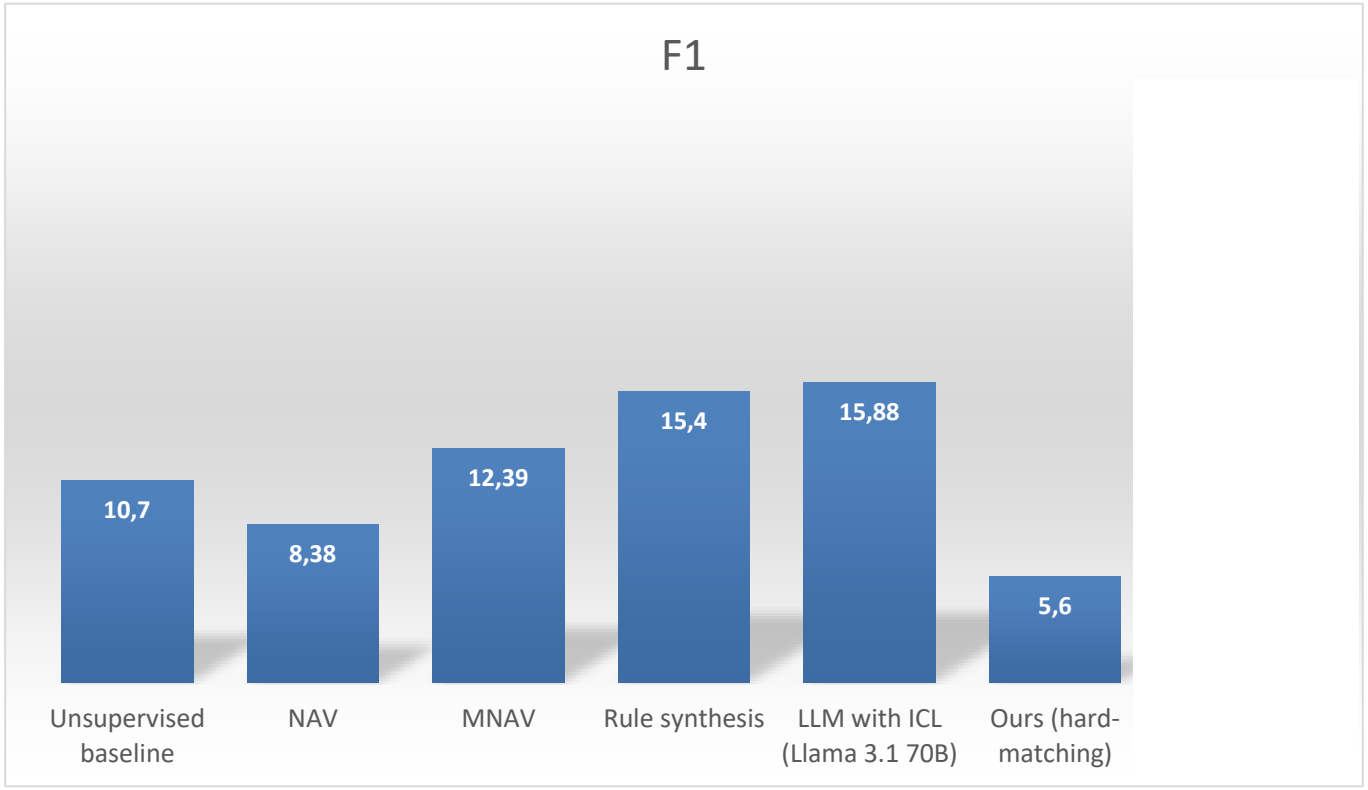
Similar to code-generation approaches such as GitHub Copilot

Results (1-shot TACRED)



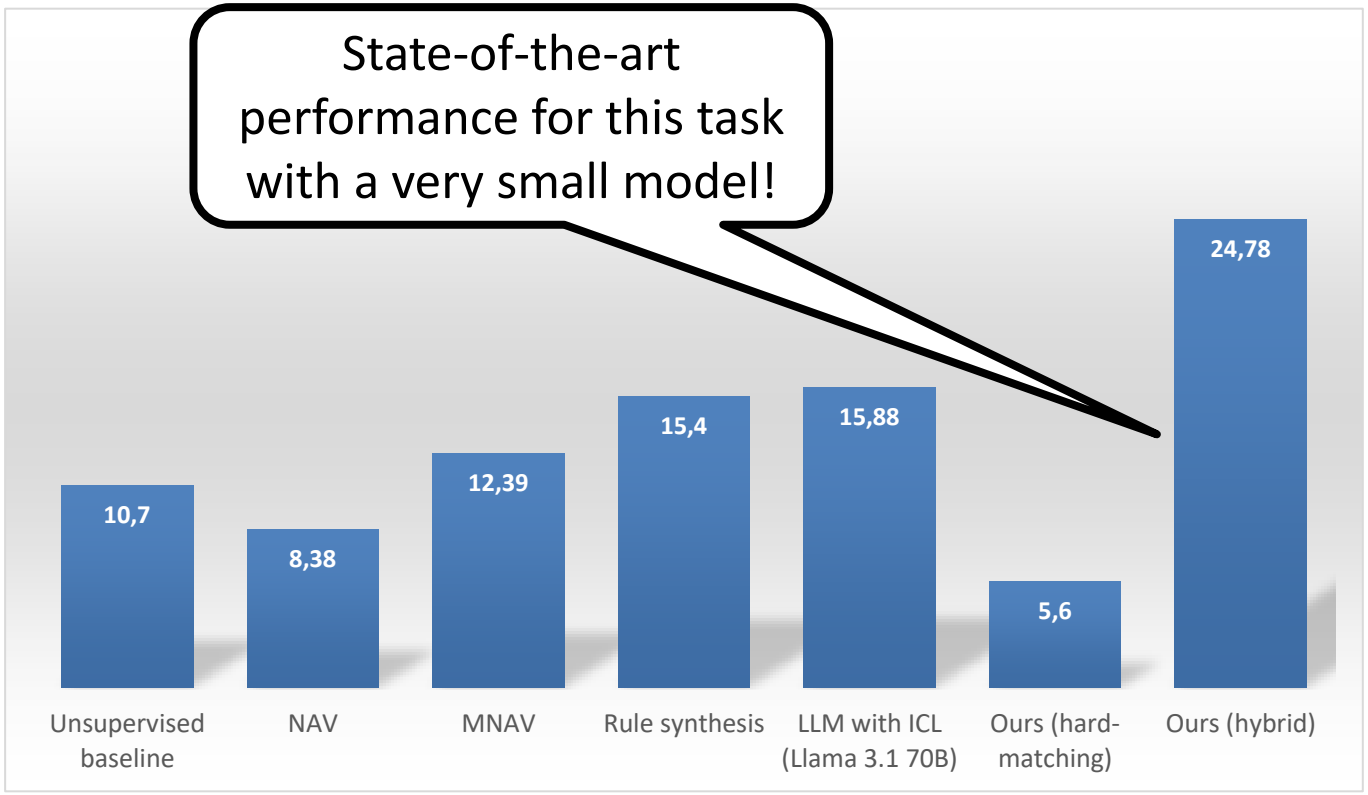
The 1-shot converted into 1 ICL example

Results (1-shot TACRED)



"Hard" rules have high precision, low recall

Results (1-shot TACRED)

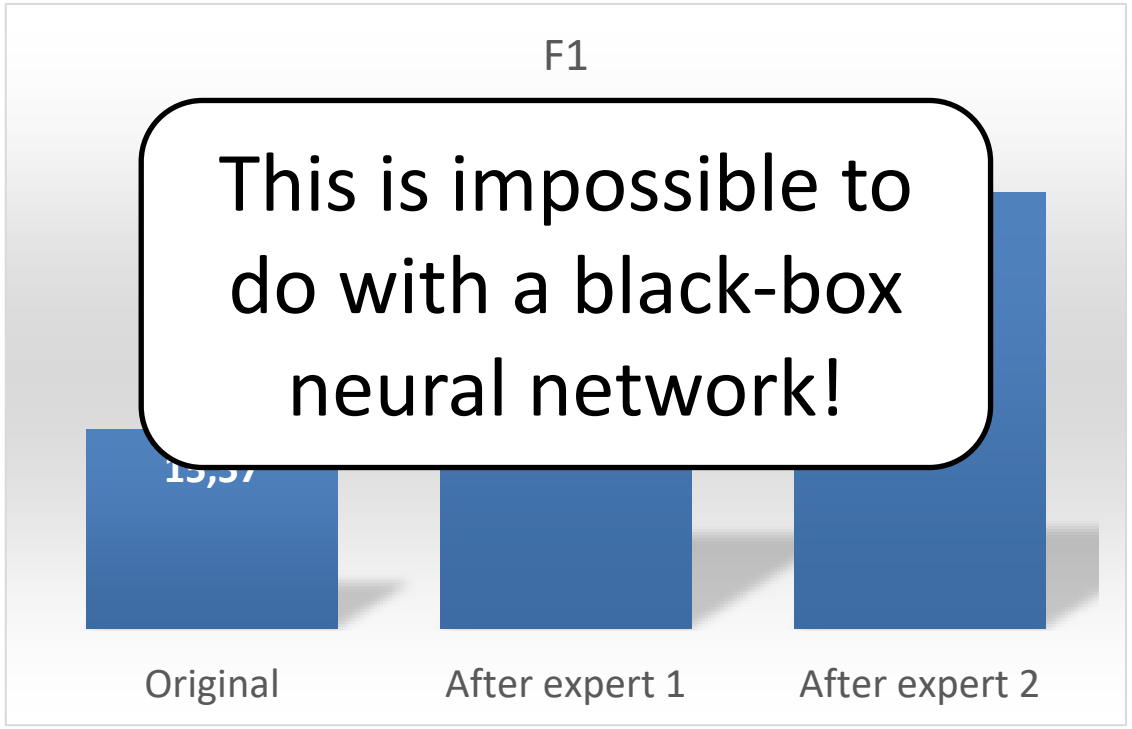


Is our approach pliable?

- Two domain experts had up to 2 hours to modify the “hard” rules for 1 relation (org:parents)
  - ADD rule
  - DELETE rule
  - MODIFY rule
- But the ”soft” matching neural component is *not* modified in any way!

Results after human intervention

5-shot setting



Conclusions so far

- A new neuro-symbolic architecture for information extraction that is modular and pliable
- Outperforms other neural networks and LLMs on a few-shot information extraction task
  - Human-in-the-loop corrections improve results further!
- Very small model
  - 1.3MB semantic matcher + a few KBs for rules

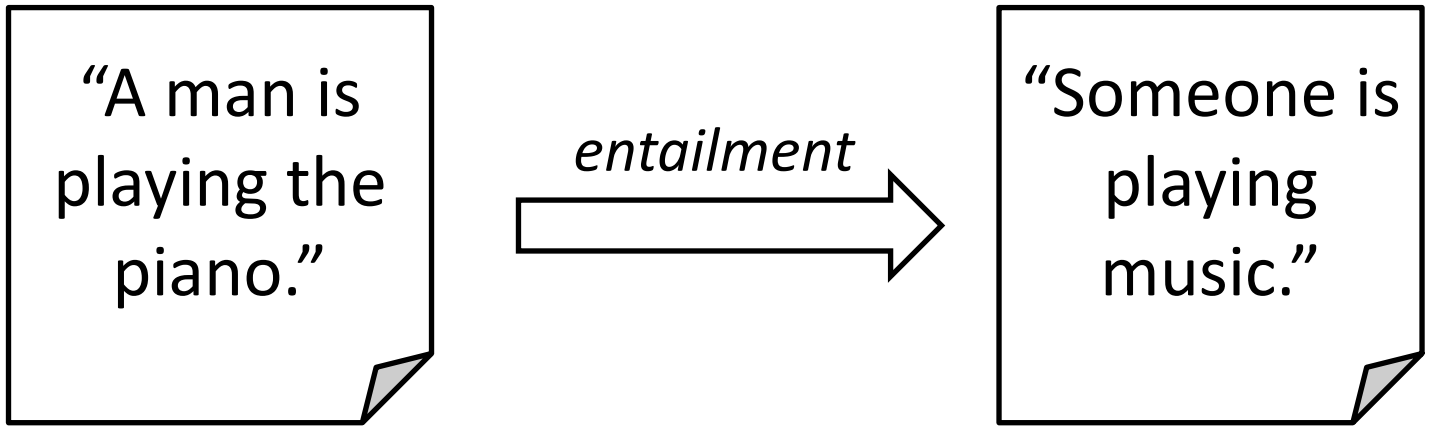
Plan

Paper #1: A neuro-symbolic approach for information extraction

Paper #2: A modular approach for natural language inference

… (talk to me offline)

What is natural language inference?



Determining if one text (*premise*) entails/contradicts/is neutral for a second text (*hypothesis*)

Why NLI is important: computational argumentation theory may improve the reasoning potential of LLMs



402 2.5k 2k 144 18.2k

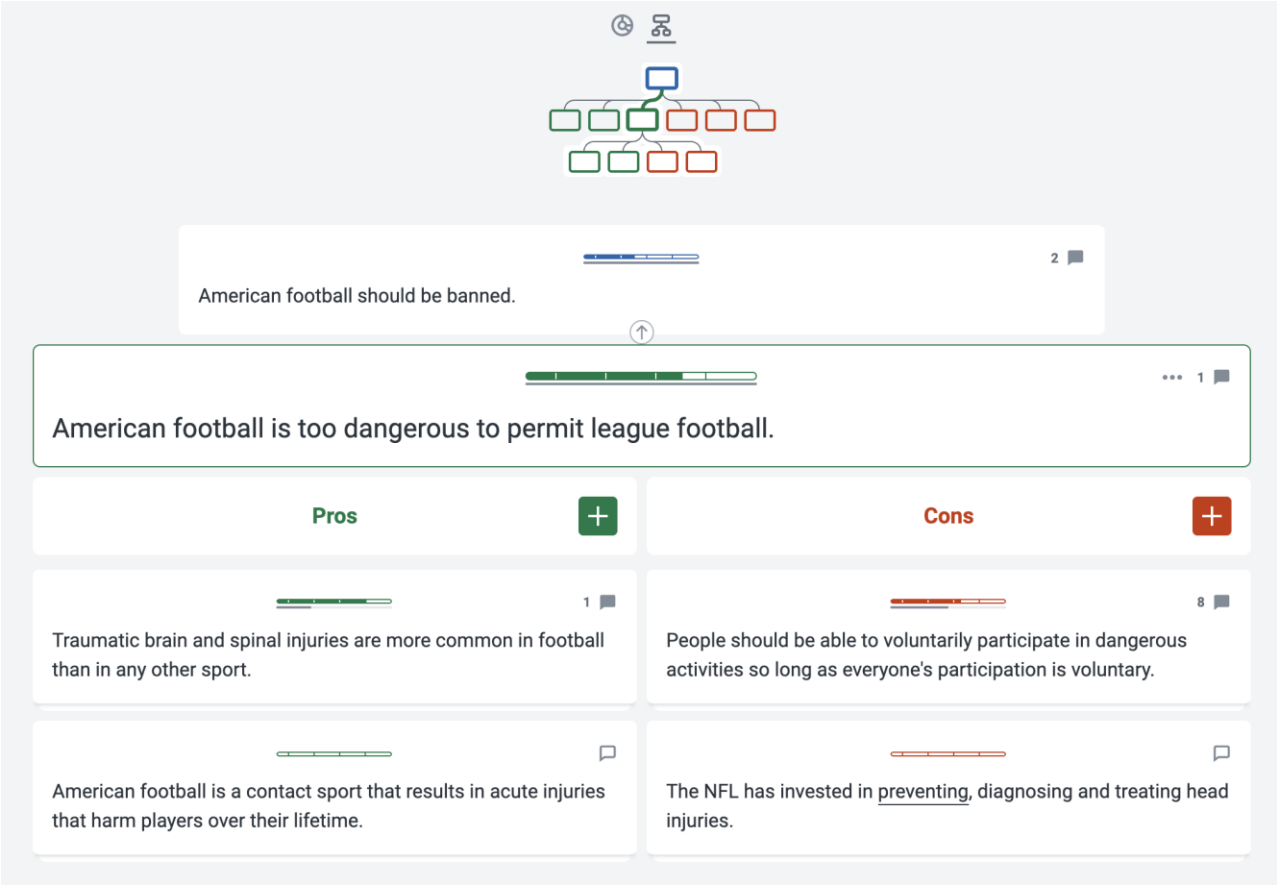
Background Info

Football is the most popular spectator sport in America but it is also one of the most controversial. Indeed, recent concerns about the safety of the game, as well as its social and economic impacts have led critics to call to ban it altogether. This debate asks whether the public play of American football should be banned at all levels.

A century ago, football was so dangerous that President Theodore Roosevelt had to step in and force reforms that ended up saving the sport. Since then, football has become an important part of America's social and economic fabric, woven into after-school activities from elementary school through high school and central to social life on many college campuses. Star football players are often among the most widely admired in their communities, and the best NFL (National Football League) players are A-list celebrities.

Example from https://www.kialo.com/should-american-football-be-banned-10143

Why NLI is important: computational argumentation theory may improve the reasoning potential of LLMs



Example from https://www.kialo.com/should-american-football-be-banned-10143

Paper #2

MorphNLI: A Stepwise Approach to Natural Language Inference Using Text Morphing

Vlad-Andrei Negru[1], Robert Vacareanu[1,2], Camelia Lemnaru[1], Mihai Surdeanu[2], Rodica Potolea[1]



Vlad-Andrei Negru
(PhD student at TUCN)

(NAACL, 2025)

Intuition from image morphing



https://ccrma.stanford.edu/~jacobliu/368Report/index.html

Text morphing vs. black-box models

Premise:	A group of children in uniforms is standing at a gate, and no one is kissing the mother.	
Hypothesis:	A crowd of people is near the water.	Gold label: neutral

Text morphing vs. black-box models

Premise: A group of children in uniforms is standing at a gate, and no one is kissing the mother.

Hypothesis: A crowd of people is near the water. Gold label: neutral

State-of-the-art encoder-decoder

Premise <SEP> Hypothesis



contradiction

Text morphing vs. black-box models

Premise: A group of children in uniforms is standing at a gate, and no one is kissing the mother.

Hypothesis: A crowd of people is near the water.

Gold label: neutral

State-of-the-art encoder-decoder

Premise <SEP> Hypothesis



contradiction

LLM Explanation

You have to provide the label and explanation for a Natural Language Inference (NLI) task ...

Premise: ...
Hypothesis: ...

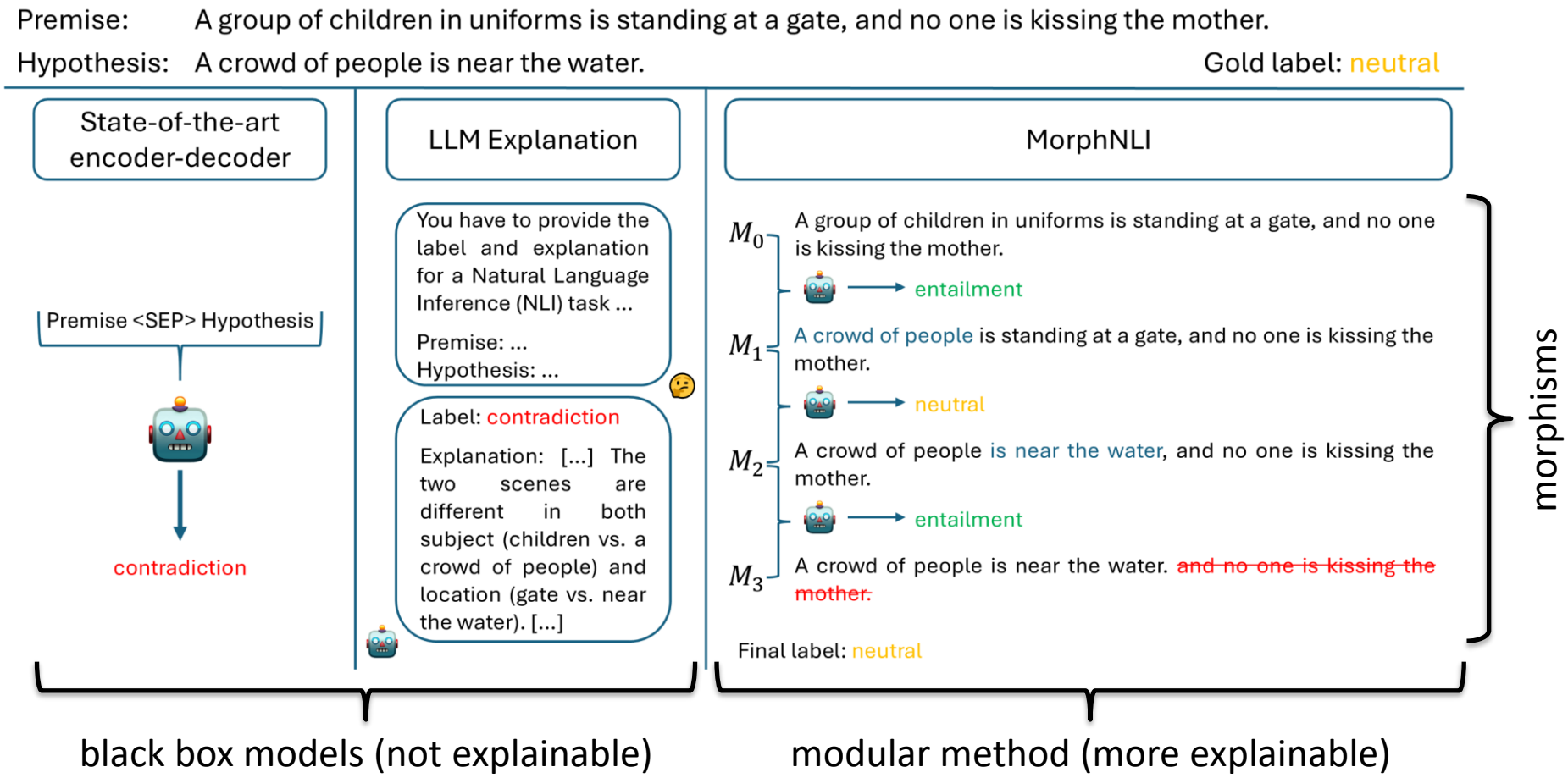


Label: contradiction

Explanation: [...] The two scenes are different in both subject (children vs. a crowd of people) and location (gate vs. near the water). [...]



Text morphing vs. black-box models

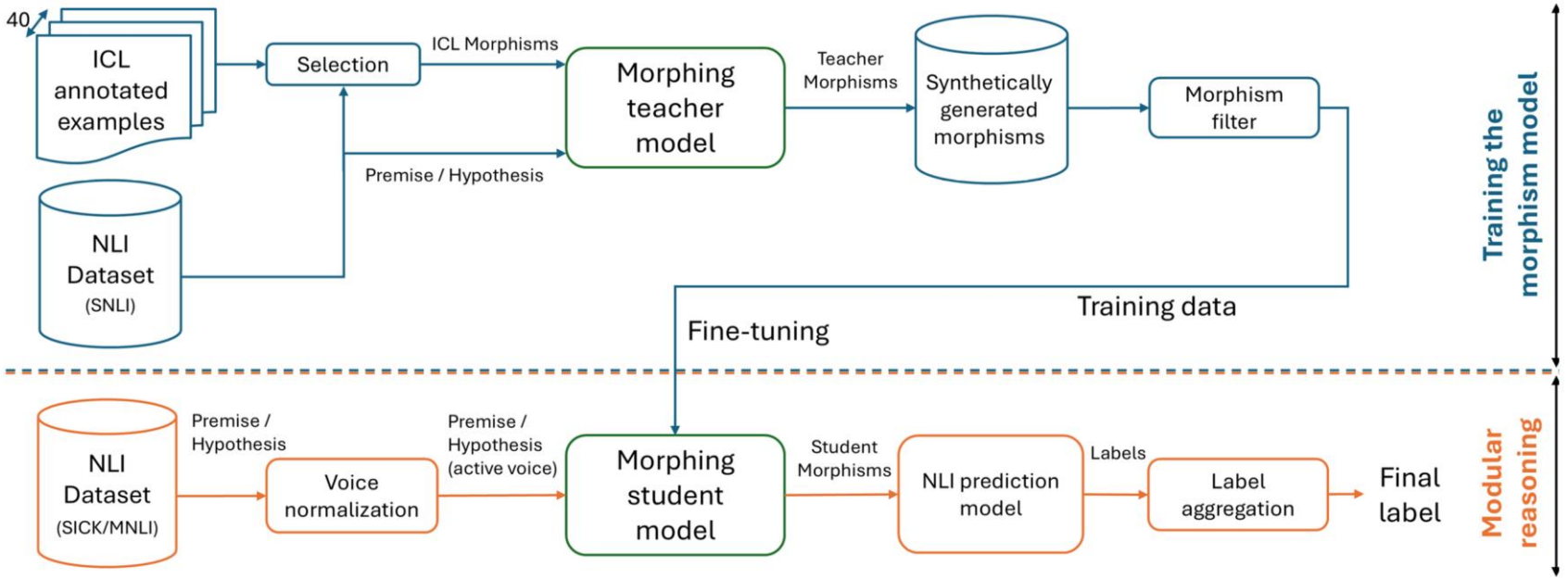


Text morphing = a pragmatic take on natural logic

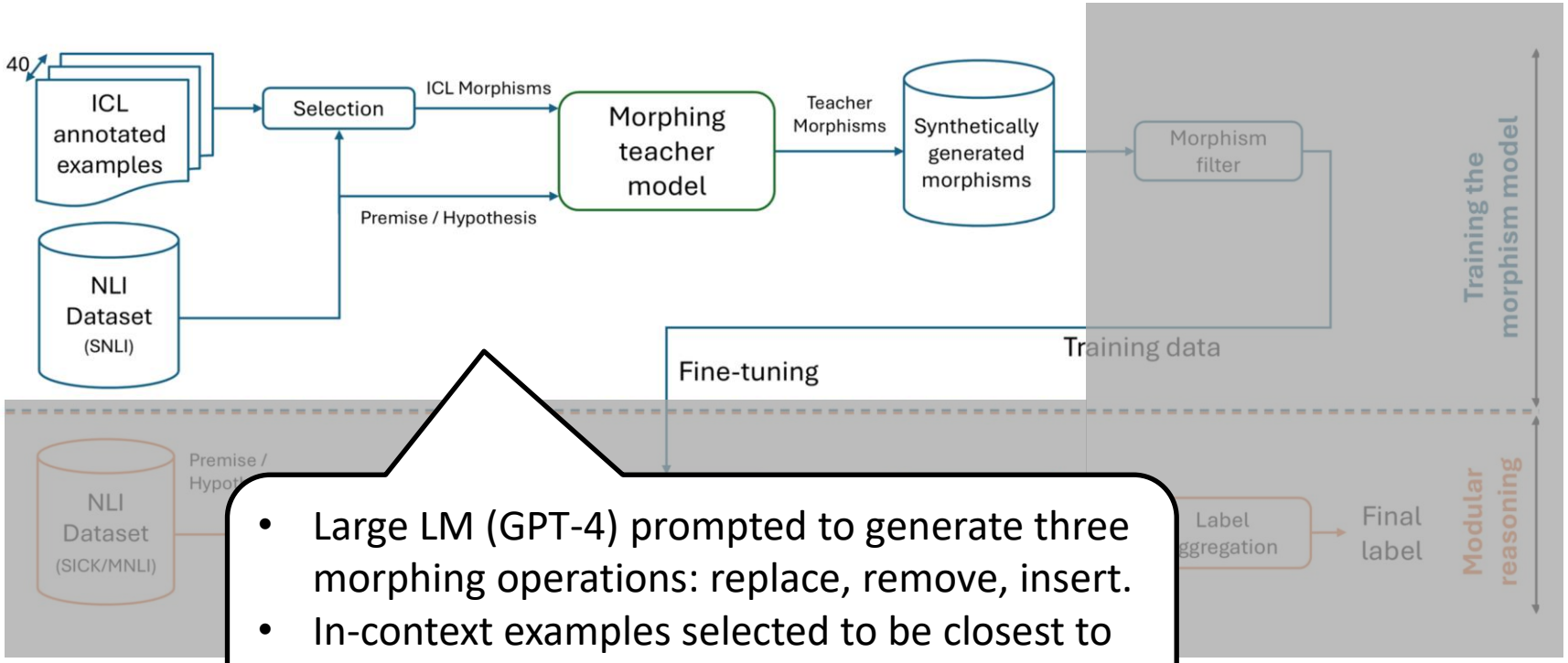
AI Overview

A natural logic example is inferring a conclusion from a statement using common-sense reasoning, such as concluding that "some sailors are giving every child some toy" from the premise "A sailor is giving a toy to each of the children", notes a presentation from Peking University. Natural logic also uses common linguistic patterns. For instance, the pattern "Every X is Y. Some Z are X. Therefore, some Z are Y" is valid, but the invalid pattern "All cats are predators. Some animals are cats. Therefore, all animals are predators" shows how a faulty conclusion is reached despite using a similar logical structure, according to a presentation from the National University of Singapore.

Approach: knowledge distillation of morphing operations

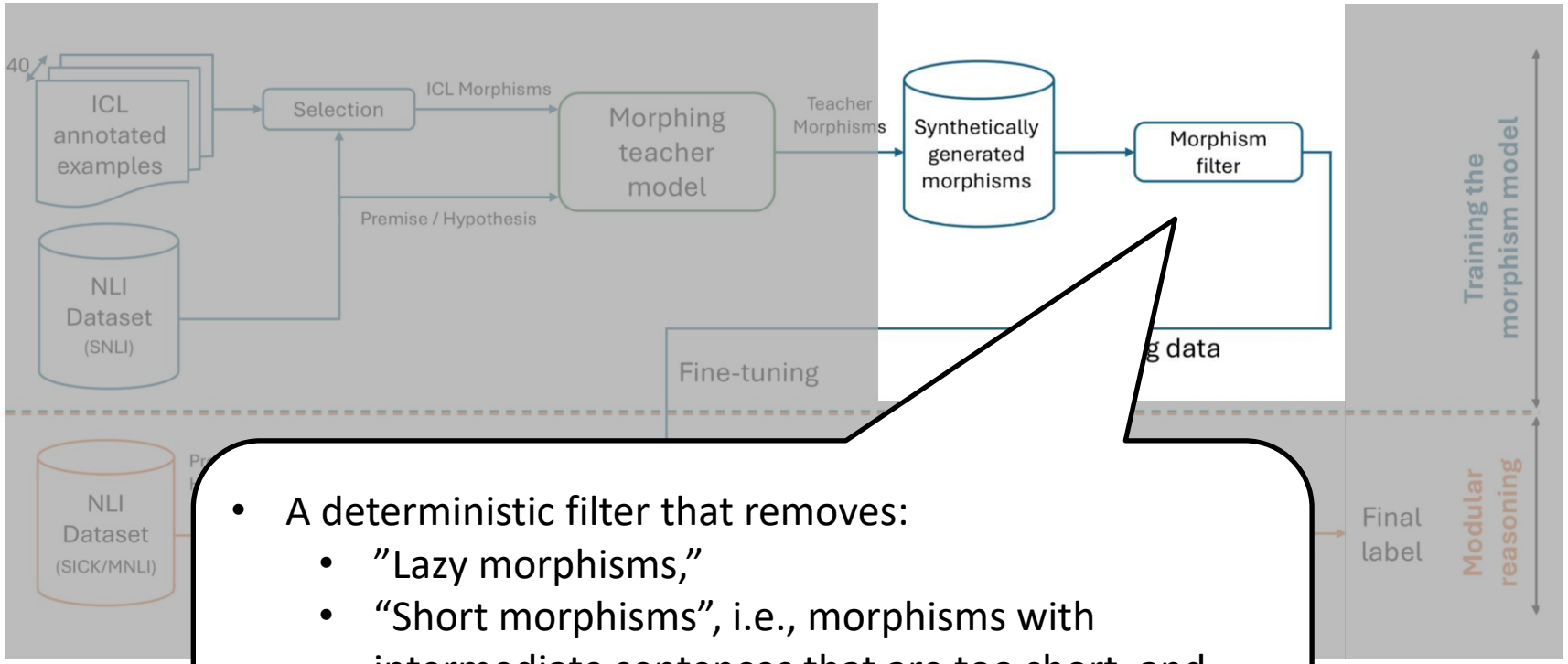


Approach: knowledge distillation of morphing operations



- Large LM (GPT-4) prompted to generate three morphing operations: replace, remove, insert.
- In-context examples selected to be closest to the data point to be morphed.

Approach: knowledge distillation of morphing operations



- A deterministic filter that removes:
  - ”Lazy morphisms,”
  - “Short morphisms”, i.e., morphisms with intermediate sentences that are too short, and
  - Morphisms that yield a different predicted overall label than gold.

Example of a short morphism

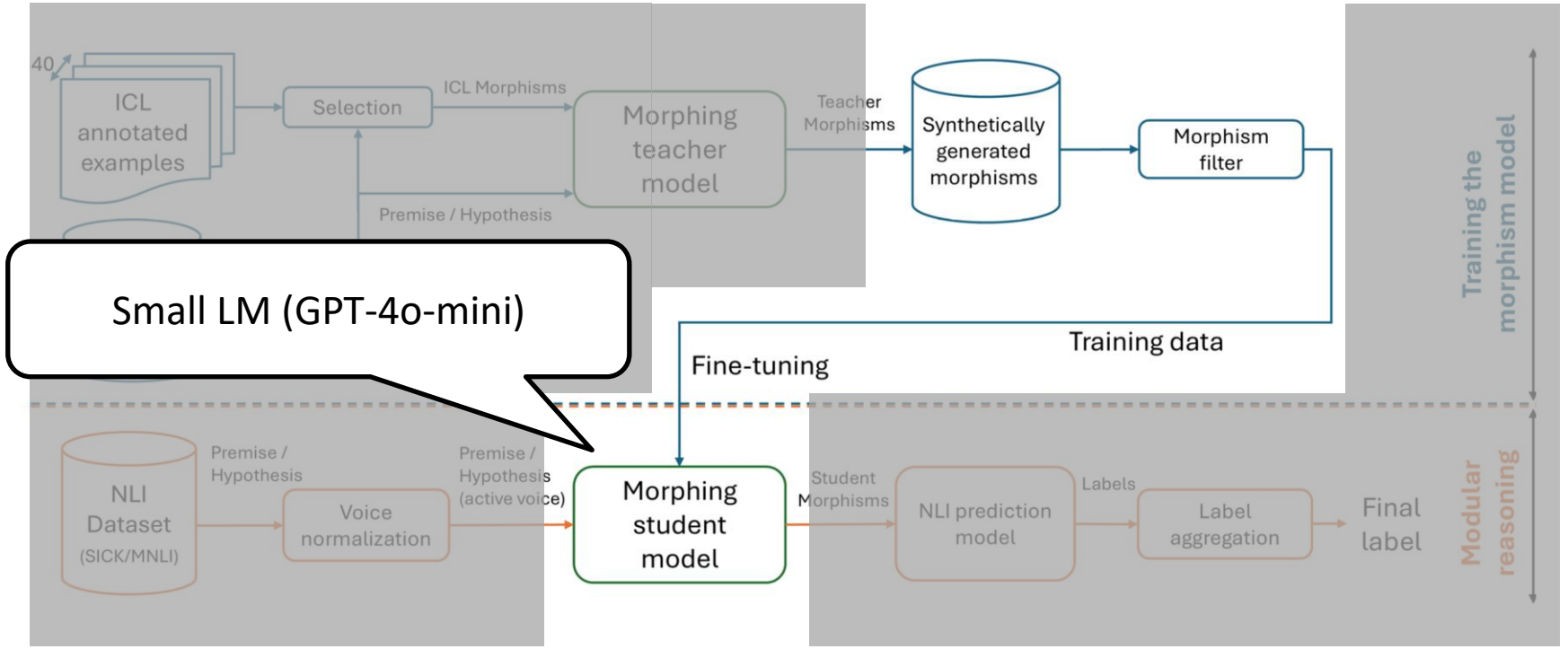
M_0: Black dog with tan markings wearing a blue collar standing on green grass.

M_1: There is a dog standing on green grass.

M_2: There is a dog standing ~~on green grass~~.

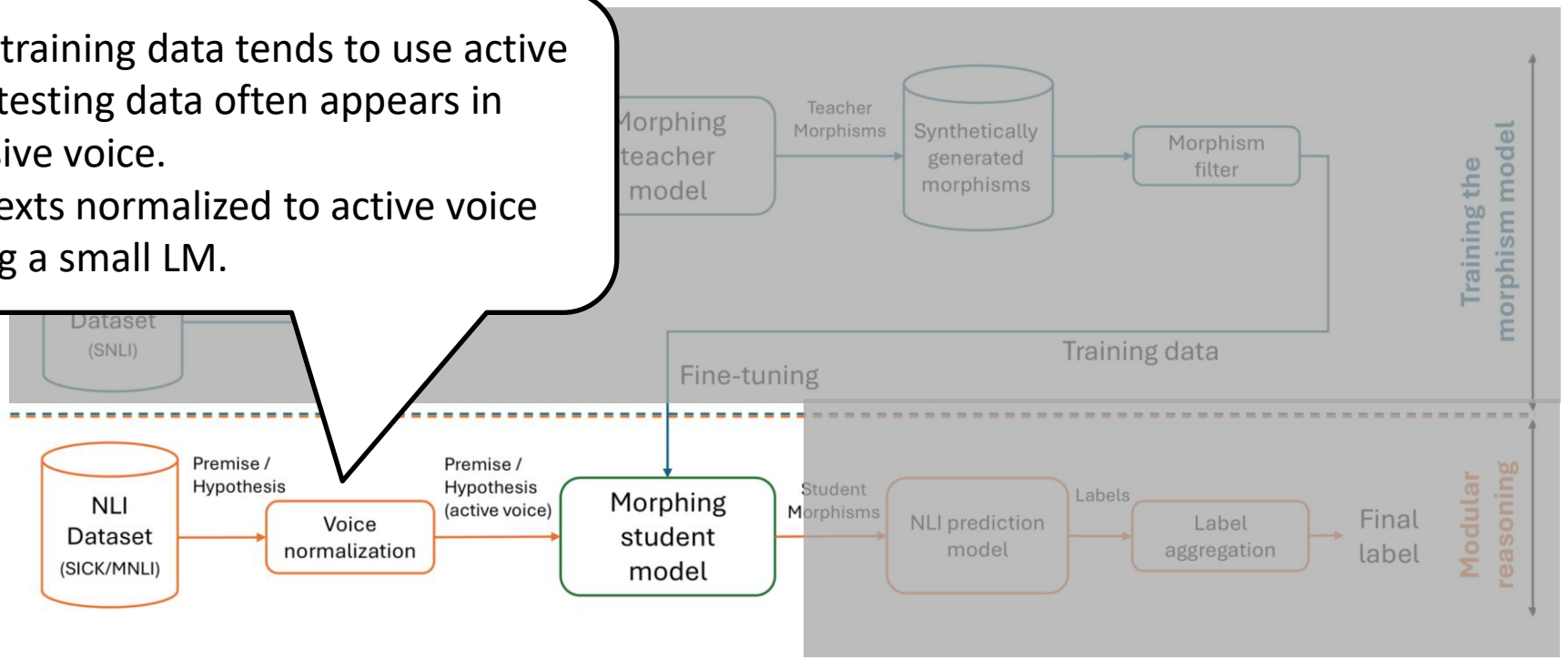
M_3: There is a dog standing outside.

Approach: knowledge distillation of morphing operations

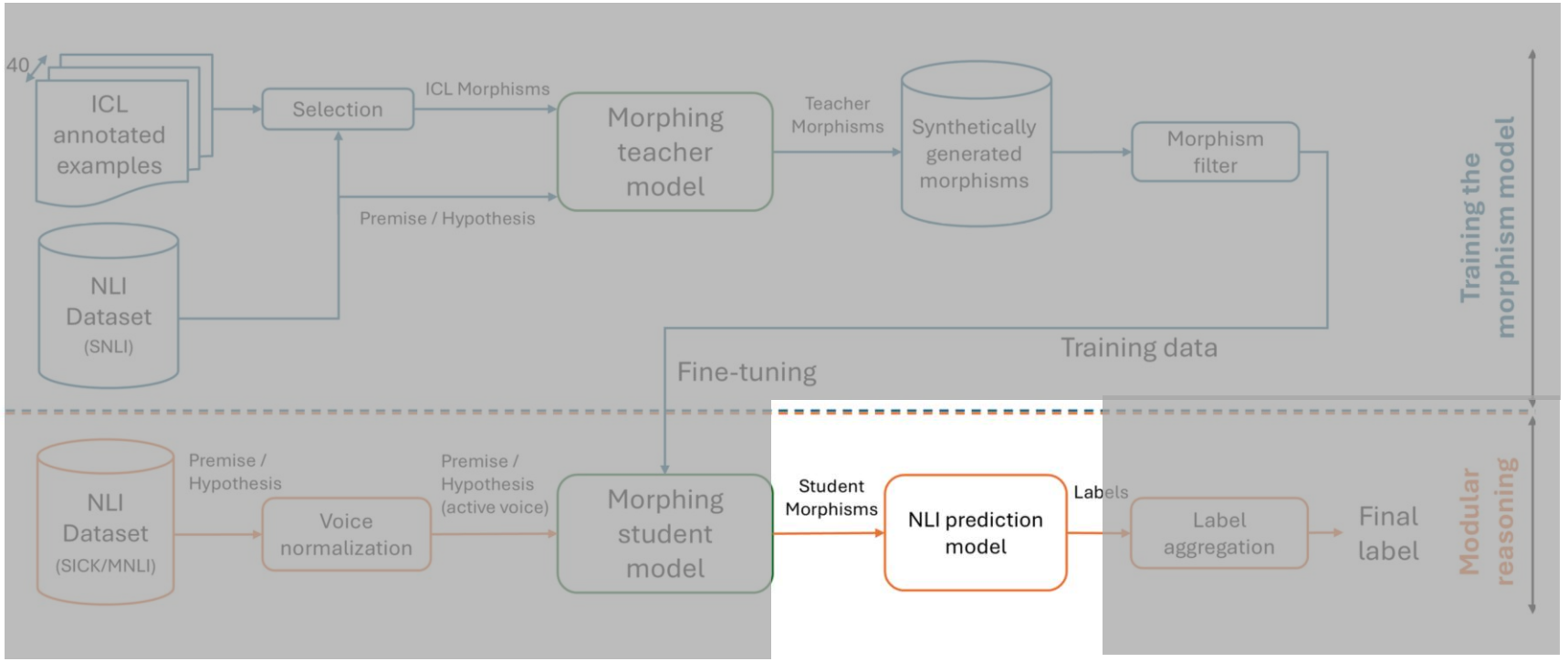


Approach: knowledge distillation of morphing operations

- The training data tends to use active but testing data often appears in passive voice.
- All texts normalized to active voice using a small LM.



Approach: knowledge distillation of morphing operations



M_0 A group of children in uniforms is standing at a gate, and no one is kissing the mother.



entailment

M_1 A crowd of people is standing at a gate, and no one is kissing the mother.



neutral

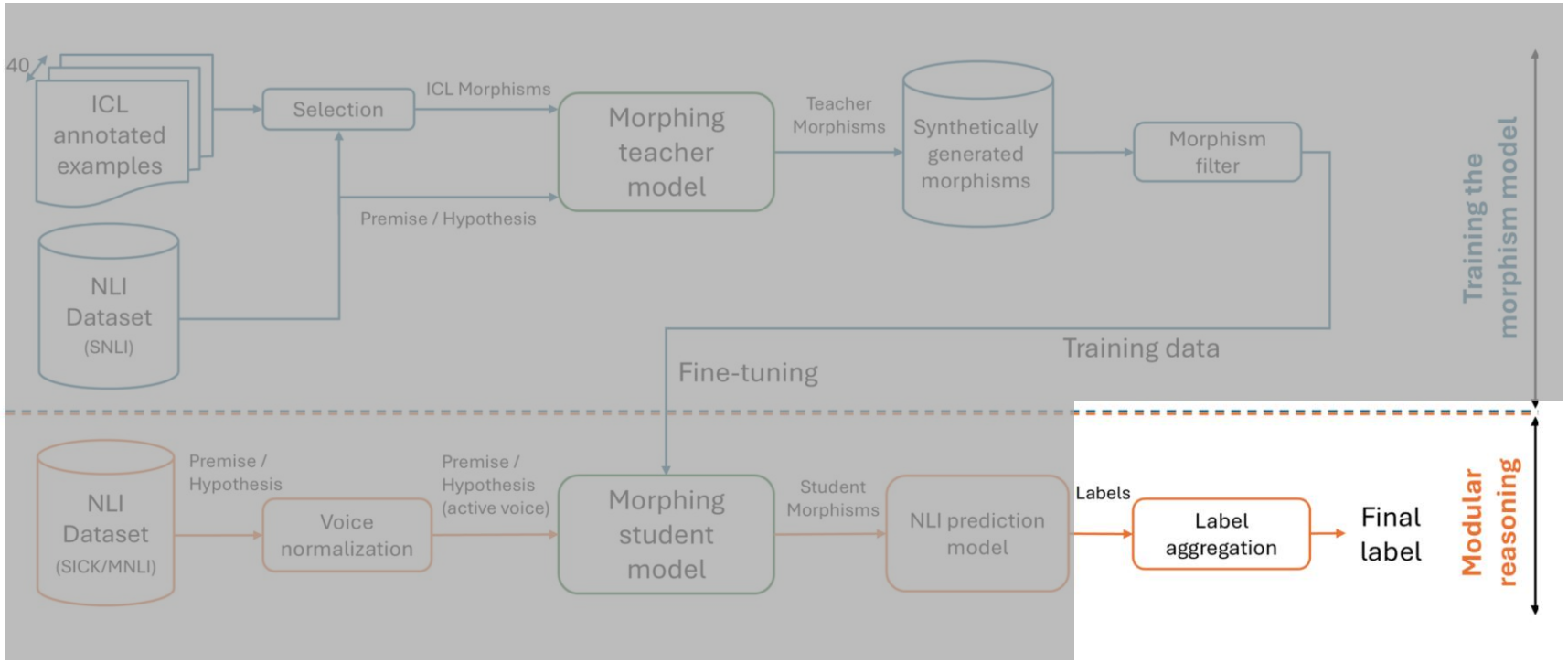
M_2 A crowd of people is near the water, and no one is kissing the mother.



entailment

M_3 A crowd of people is near the water. ~~and no one is kissing the mother.~~

Approach: knowledge distillation of morphing operations



M_0 A group of children in uniforms is standing at a gate, and no one is kissing the mother.

🤖 → entailment

M_1 A crowd of people is standing at a gate, and no one is kissing the mother.

🤖 → neutral

M_2 A crowd of people is near the water, and no one is kissing the mother.

🤖 → entailment

M_3 A crowd of people is near the water. ~~and no one is kissing the mother.~~

neutral

(first non-entailment label)

How well does it work?

- One training dataset (for morphisms): SNLI
- Two different evaluation datasets:
  - Sentences Involving Compositional Knowledge (SICK) (Marelli et al., 2014)
  - Multi-Genre Natural Language Inference (MNLI) (Williams et al., 2018)

Performance on MNLI (ID and OOD)

MNLI	ID	OOD
RoBERTa Vanilla	**89.91**	53.00
RoBERTa Vanilla (+VN)	88.50	52.77
RoBERTa Morphism	85.01	**58.29**
RoBERTa Morphism (+VN)	83.32	56.73
BART Vanilla	**88.24**	46.86
BART Vanilla (+VN)	86.48	45.50
BART Morphism	82.00	**52.78**
BART Morphism (+VN)	80.16	51.12

Performance on MNLI (ID and OOD)

MNLI		
RoBERTa Vanilla		
RoBERTa Vanilla (+VN)		
RoBERTa Morphism		
RoBERTa Morphism (+VN)		
BART Vanilla	**88.24**	46.86
BART Vanilla (+VN)	86.48	45.50
BART Morphism	82.00	**52.78**
BART Morphism (+VN)	80.16	51.12

- Blackbox models
- RoBERTa is encoder-only
- BART is encoder-decoder

Performance on MNLI (ID and OOD)

MN		OOD
RoB		53.00
RoB		52.77
RoB		**58.29**
RoB	2	56.73
BART Vanilla	**88.24**	46.86
BART Vanilla (+VN)	86.48	45.50
BART Morphism	82.00	**52.78**
BART Morphism (+VN)	80.16	51.12

The only evaluation that should matter is out of domain!

Our morphing-based method performs considerably better OOD

(ID and OOD)

	ID	OOD
RoBERTa Vanilla	**89.91**	53.00
RoBERTa Vanilla (+VN)	[illegible]	52.77
RoBERTa Morphism	85.01	**58.29**
RoBERTa Morphism (+VN)	83.32	56.73
BART Vanilla	**88.24**	46.86
BART Vanilla (+VN)	86.48	45.50
BART Morphism	82.00	**52.78**
BART Morphism (+VN)	80.16	51.12

Performance on SICK (ID and OOD)

SICK	ID	OOD
RoBERTa Vanilla	90.64	56.62
RoBERTa Vanilla (+VN)	**90.91**	56.52
RoBERTa Morphism	88.14	57.68
RoBERTa Morphism (+VN)	88.32	**57.94**
BART Vanilla	89.85	59.29
BART Vanilla (+VN)	**90.07**	58.64
BART Morphism	87.38	59.64
BART Morphism (+VN)	88.59	**60.38**

Are the explanations good?

Model	SICK	MNLI
MorphNLI explanations	**70.63**	70.00
GPT-4o explanations	62.50	**92.50**
Llama 3.1 8B explanations	47.50	56.67

ns good?

GPT models have been contaminated with the MNLI dataset!

Model	SICK	MNLI
MorphNLI explanations	70.63	70.00
GPT-4o explanations	62.50	**92.50**
Llama 3.1 8B explanations	47.50	56.67

Our explanations (green) have higher quality than explanations coming from much larger models (red)!

nations good?

	SICK	MNLI
	70.63	70.00
GPT-4o explanations	62.50	**92.50**
Llama 3.1 8B explanations	47.50	56.67

Most common errors

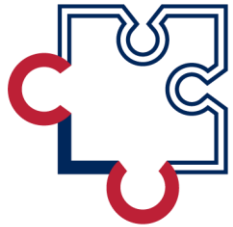
	SICK	MNLI
NLI model errors	45%	50%
Morphing errors	45%	20%
Voice normalization errors	5%	30%

Conclusions

- LLMs are great but their drawbacks limit their applicability in critical domains.
- This talk introduced other architectures for natural language processing that are fast, explainable, pliable, and perform well!

THANK YOU!

Emails: msurdeanu@arizona.edu



https://clulab.org