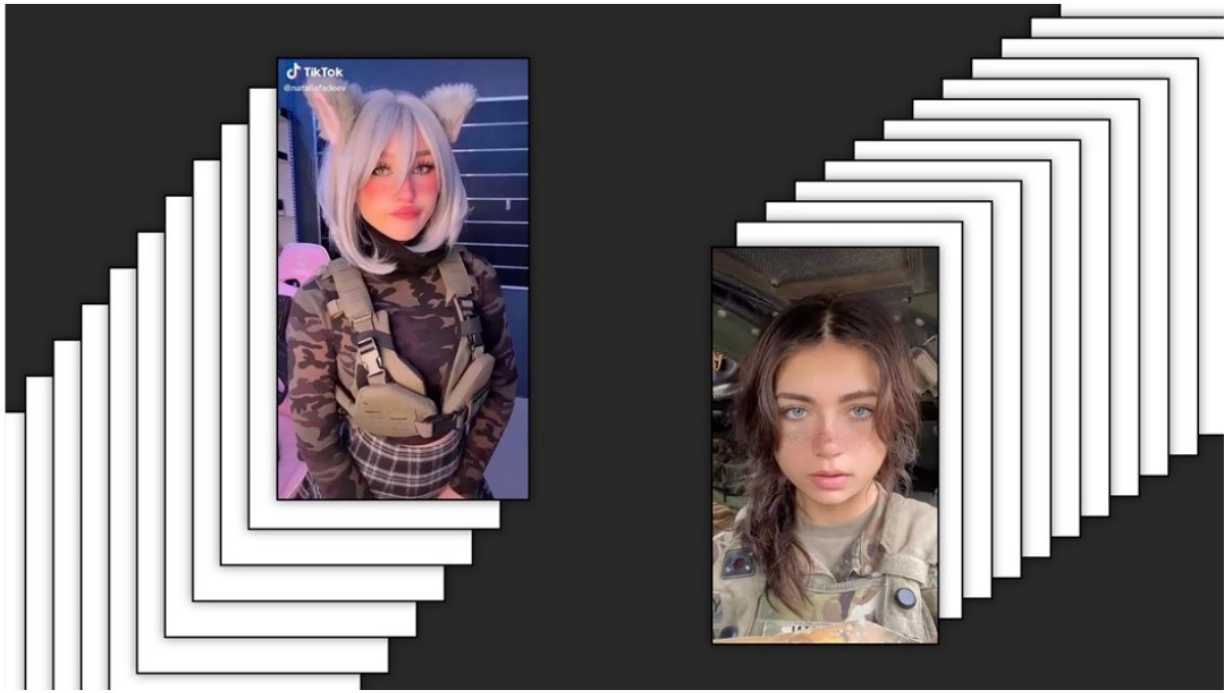
Fundamentals of Large Generative Language Models (LLMs) for TM

Maxime Würsch, EPFL & CYD Campus
Andrei Kucharavy, HES-SO Wallis/Valais
Dimitri Percia-David, HES-SO Wallis/Valais

The New York Times

A.I. Is Becoming More Conversational. But Will It Get More Honest?

At a new website called Character.AI, you can chat with a reasonable facsimile of almost anyone, live or dead, real or (especially) imagined.



How E-girl influencers are trying to get Gen Z into the military

Cosplay commandos are posting nationalist thirst traps to mobilise the SIMPs – but why?

Featured Article

Is ChatGPT a cybersecurity threat?

Carly Page @carlypage_ / 9:30 PM GMT+1 • January 11, 2023

Comment

Forbes

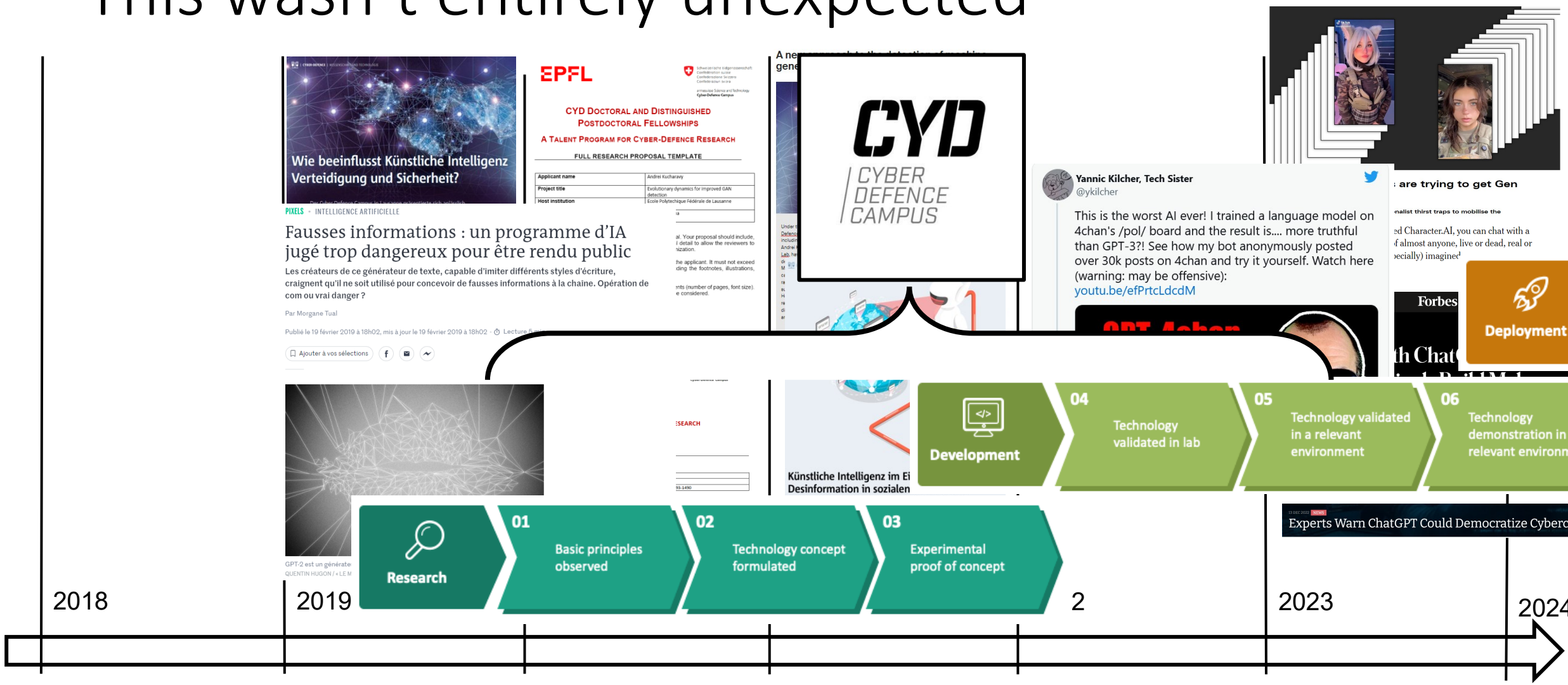
CYBERSECURITY • EDITORS' PICK

Armed With ChatGPT, Cybercriminals Build Malware And Plot Fake Girl Bots

13 DEC 2022 NEWS

Experts Warn ChatGPT Could Democratize Cybercrime

This wasn’t entirely unexpected



Indirect technological impact?

Fundamentals of Generative Large Language Models and Perspectives in Cyber-Defense

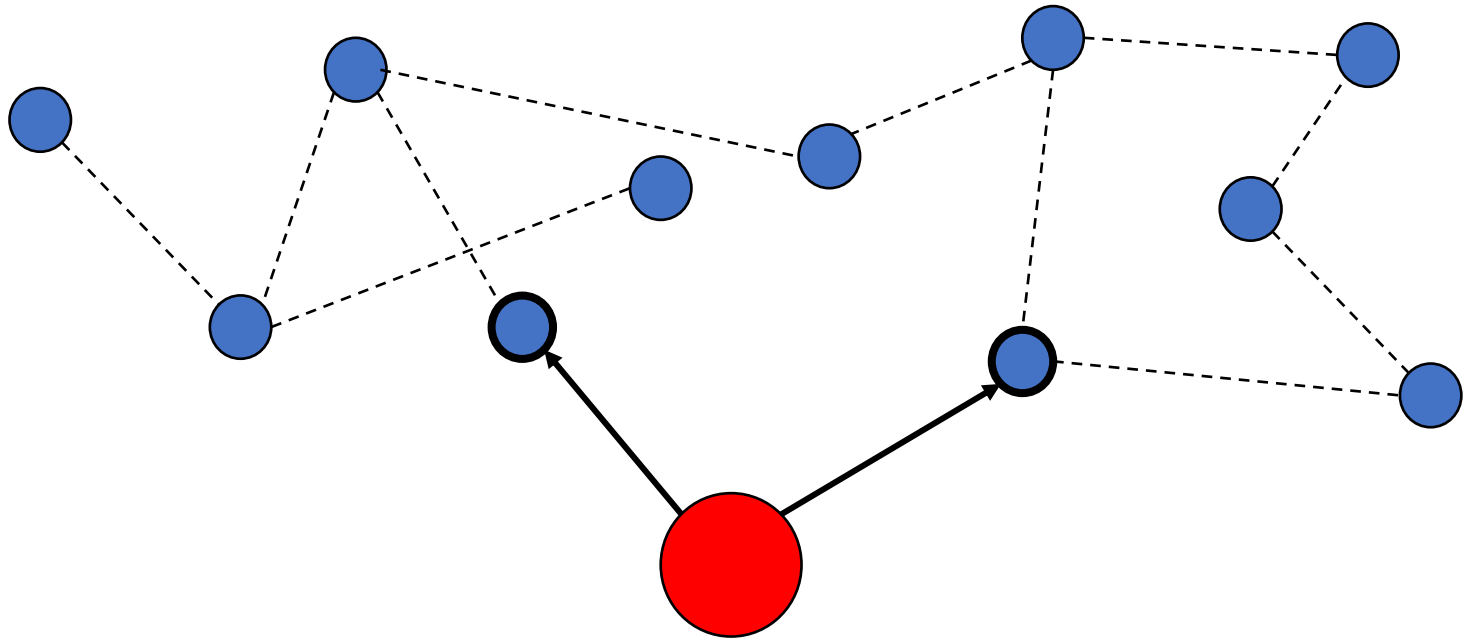
Andrei Kucharavy[1, 2, +], Zachary Schillaci[3], Loïc Maréchal[1,4], Maxime Würsch[1,5], Ljiljana Dolamic[1], Remi Sabonnadiere[3], Dimitri Percia David[1,2], Alain Mermoud[1], and Vincent Lenders[1]

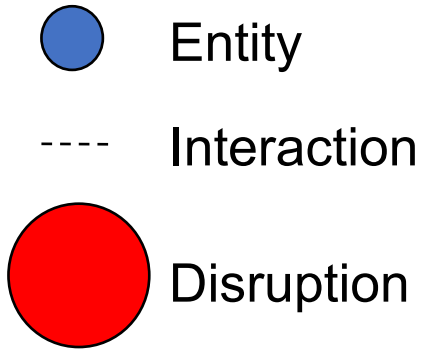
[+] *Corresponding Author; andrei.kucharavy@hevs.ch*
[1] *Cyber-Defence Campus, armasuisse S+T*
[2] *Institute of Entrepreneurship & Management, HES-SO Valais-Wallis*
[3] *Effixis SA*
[4] *Department of Information Systems, HEC Lausanne, University of Lausanne*
[5] *Section of Computer Science, EPFL*

https://arxiv.org/abs/2303.12132

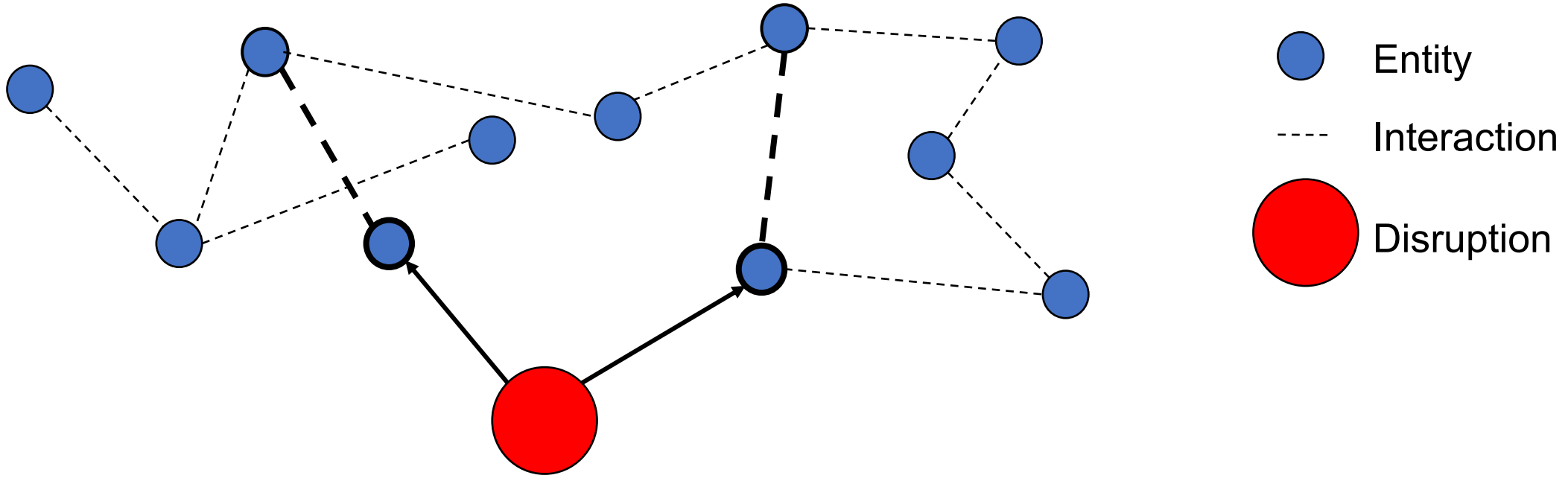
(March 2022)

Can we do better?

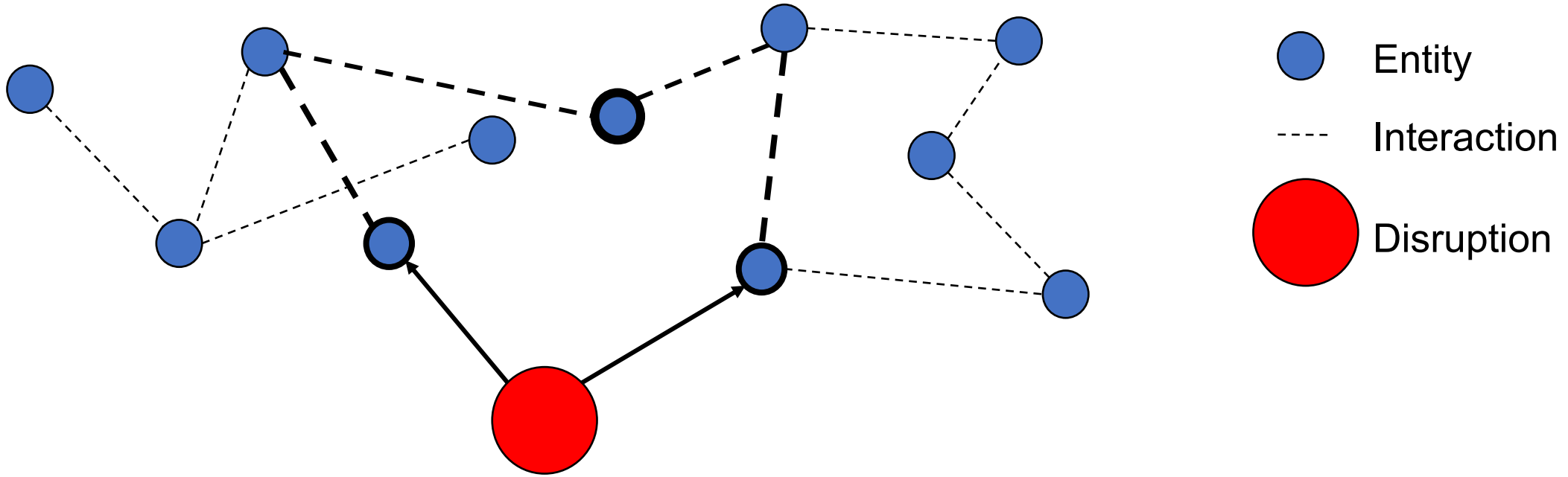




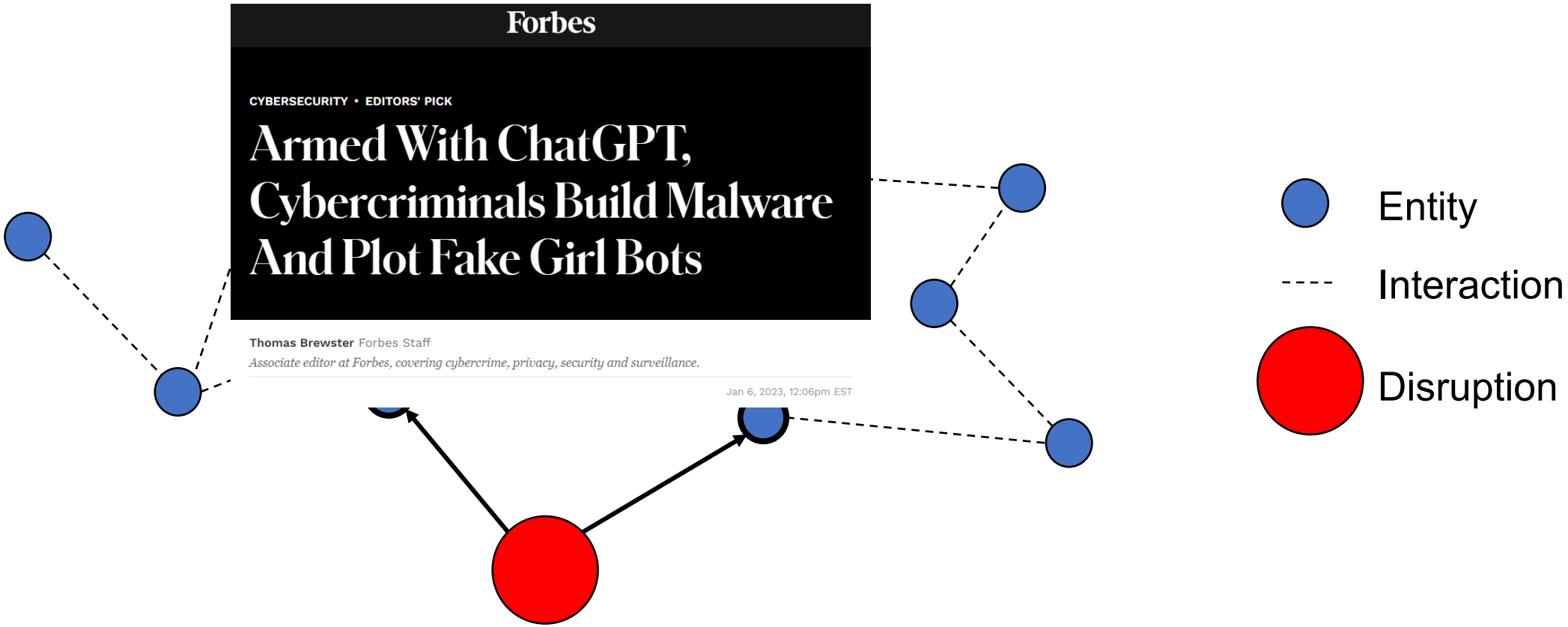
Can we do better?



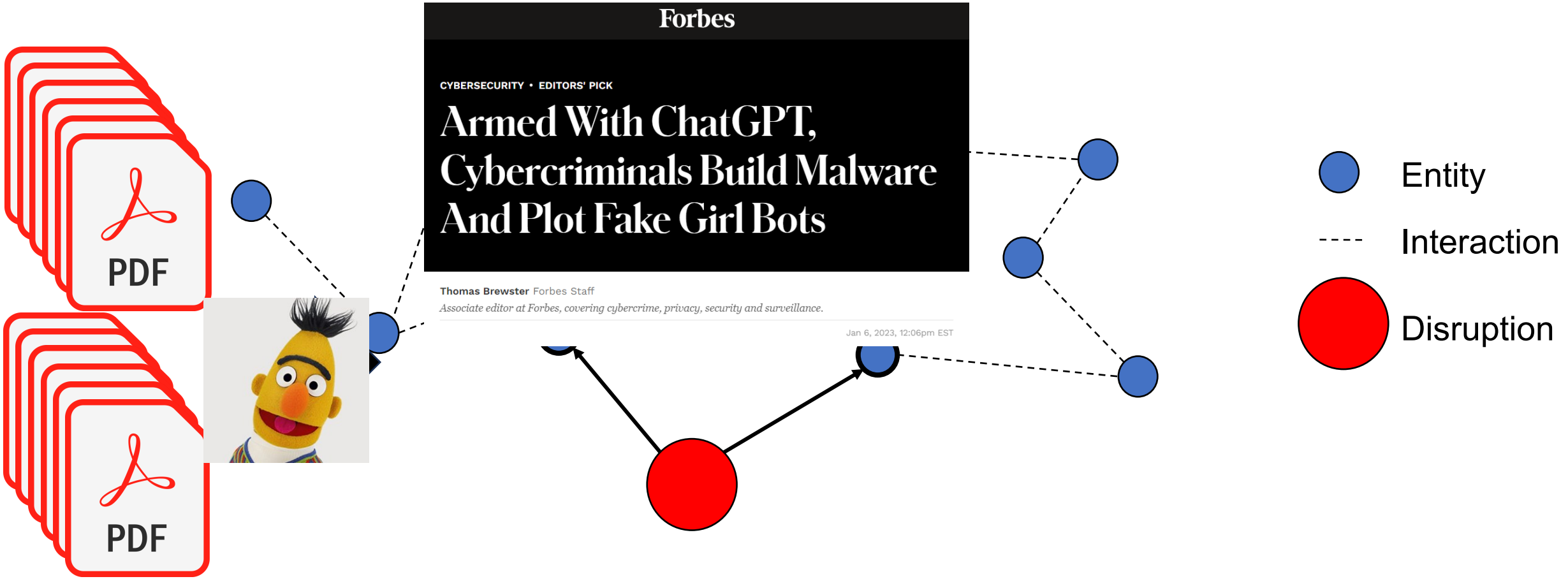
Can we do better?



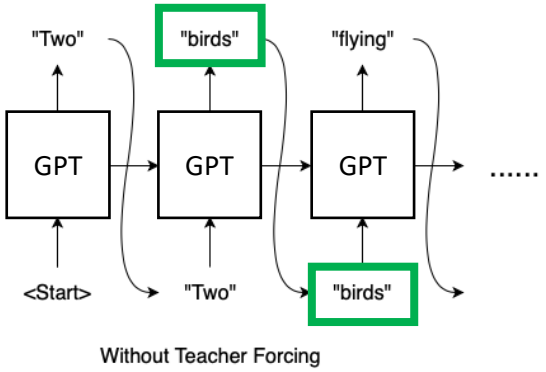
Can we do better?



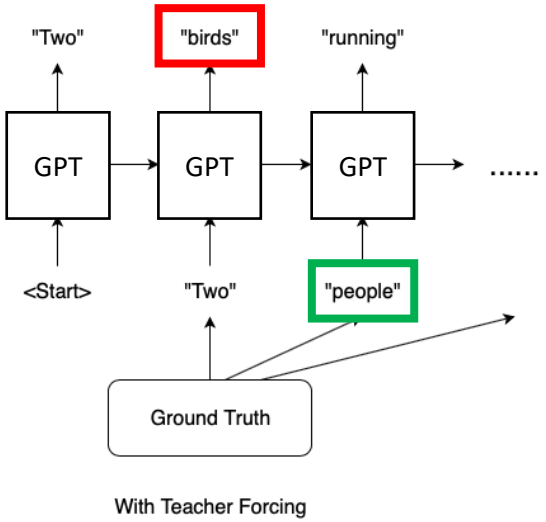
Can we do better?



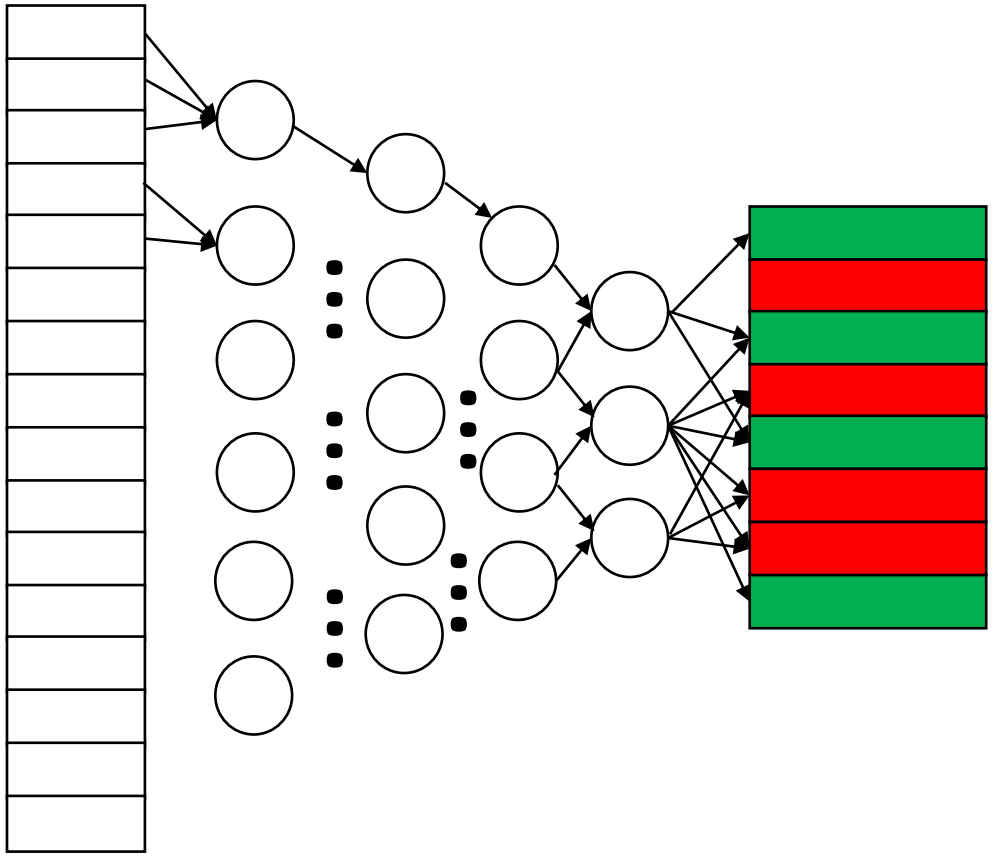
What LLMs “Learn”



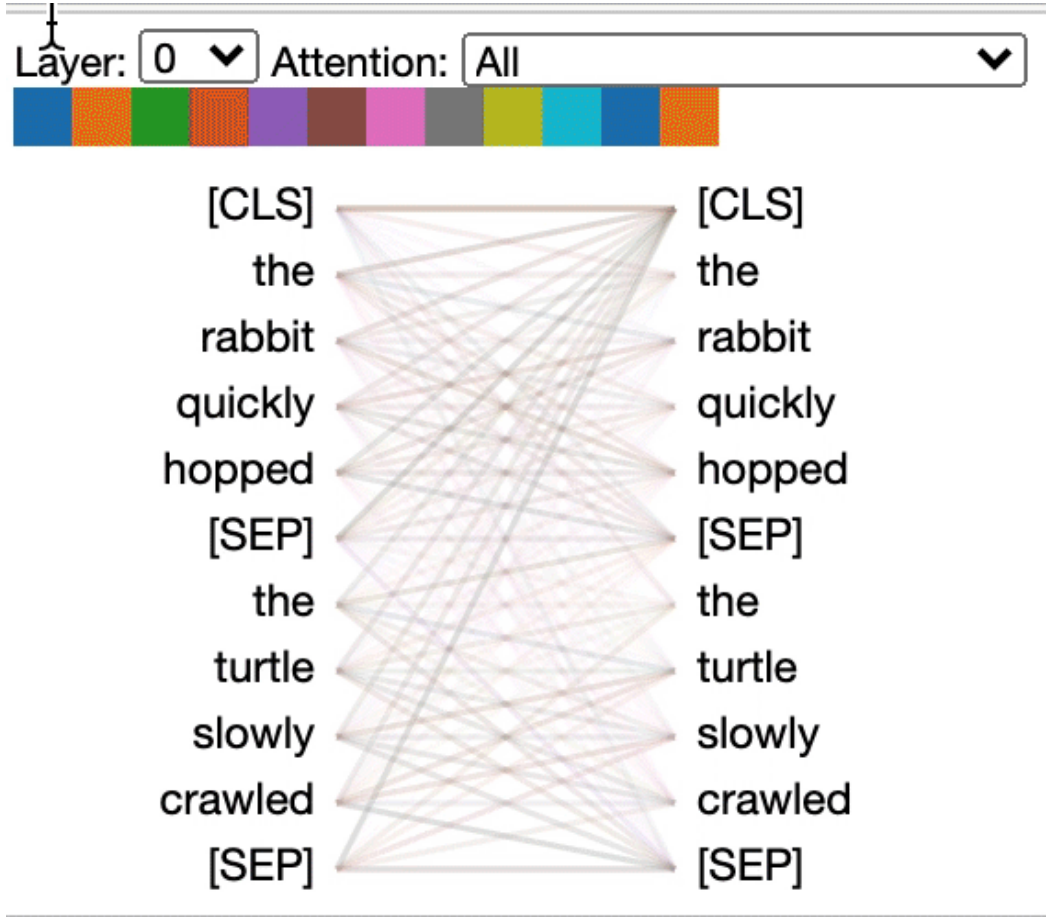
Without Teacher Forcing

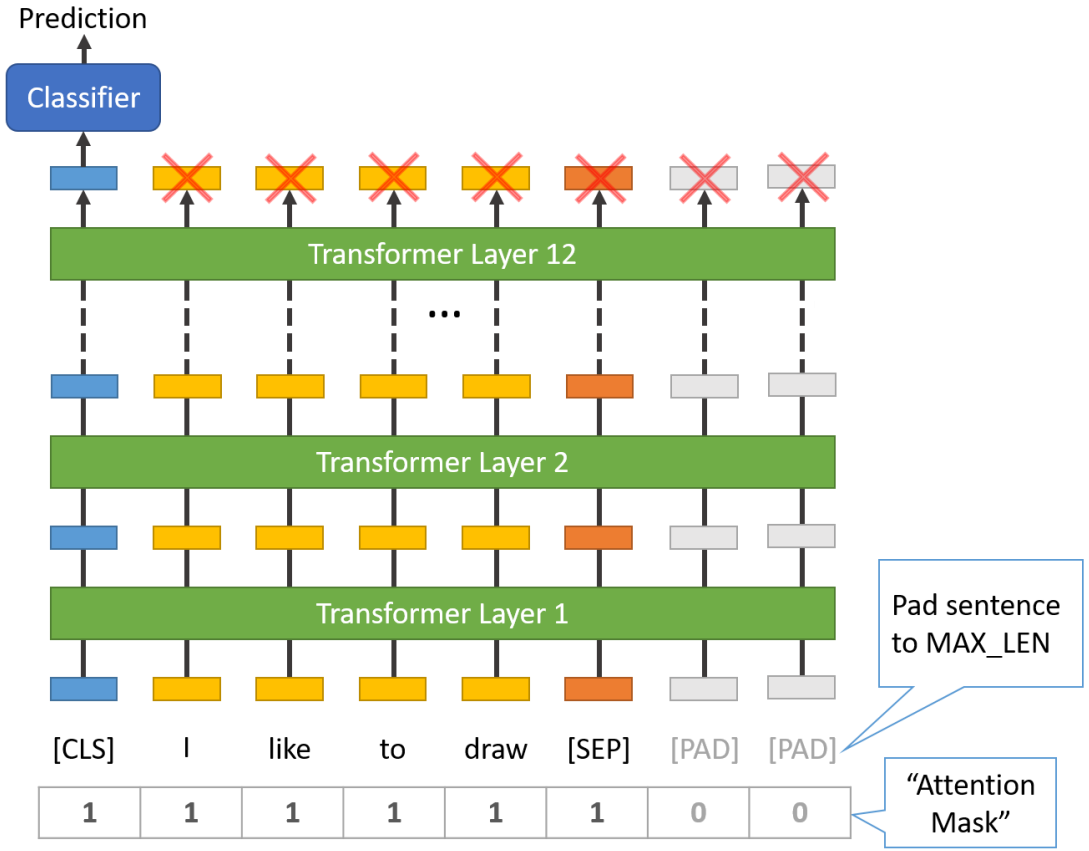


With Teacher Forcing

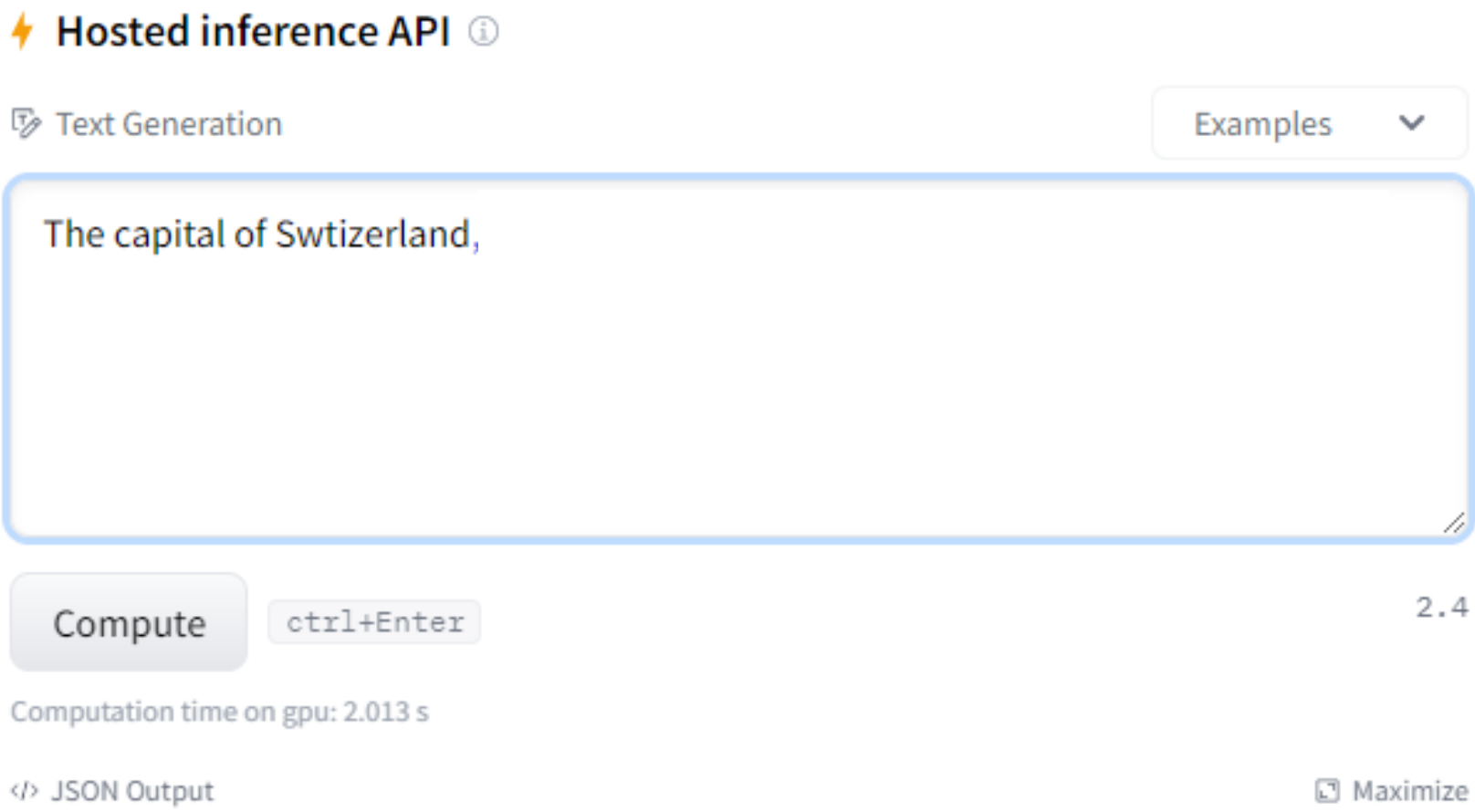


Autoencoder Fine-Tunes





LLMs have Failure Modes (Due to the Training Data)



GPT-neo-2.7B

(And some Have Implications)

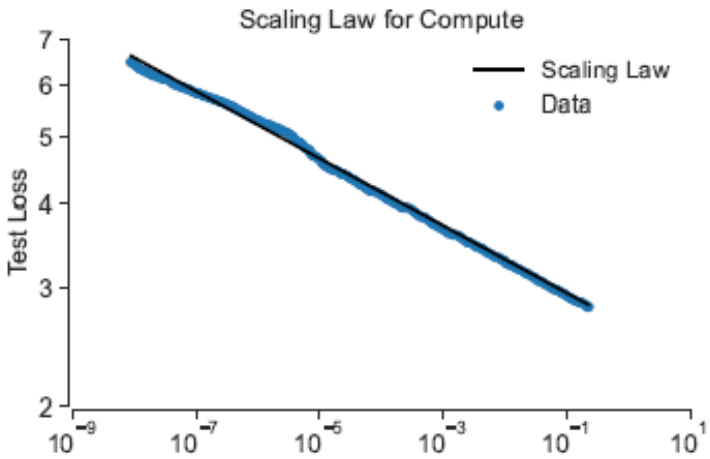
Flan T5-UL2

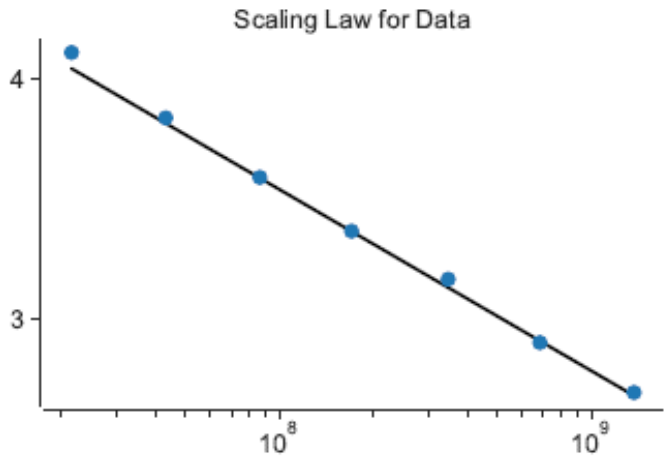
The Swiss Army has a very small military. The Swiss Army h
and require a large amount of manpower to operate. There

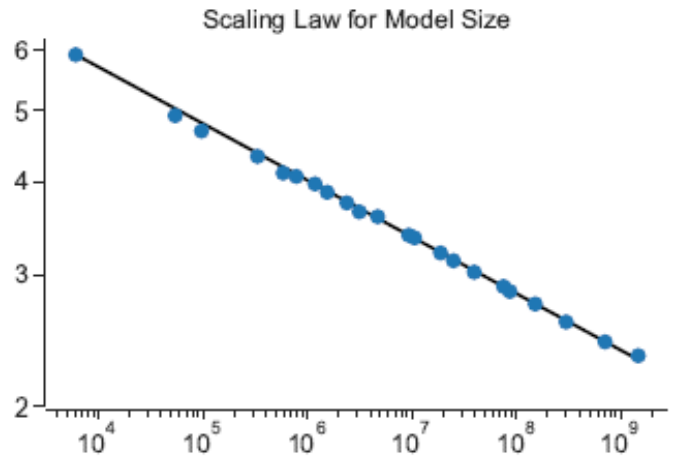
Flan T5-XXL

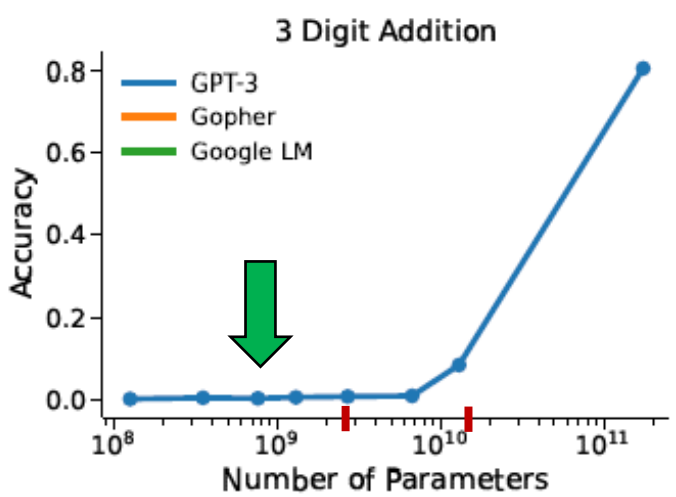
The Swiss Army is a land force. Land forces do not use tank

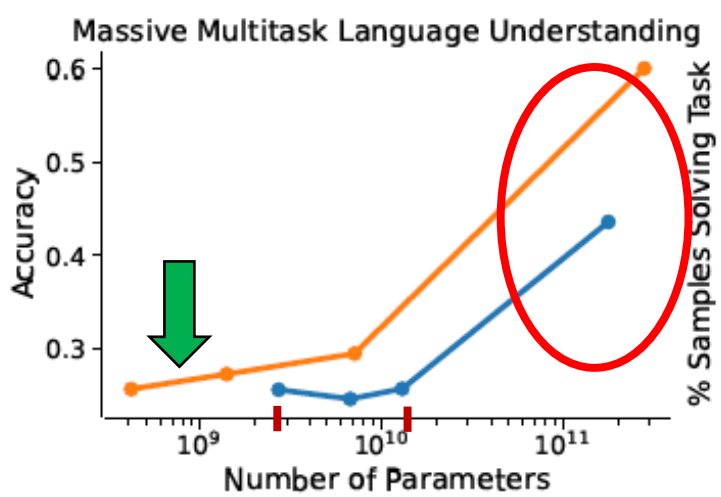
Size Matters, but
Smaller LLMs Provide Insight on Larger Ones

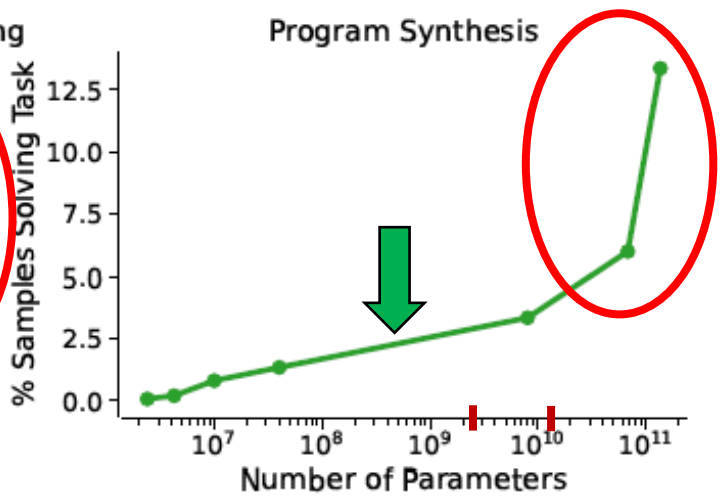




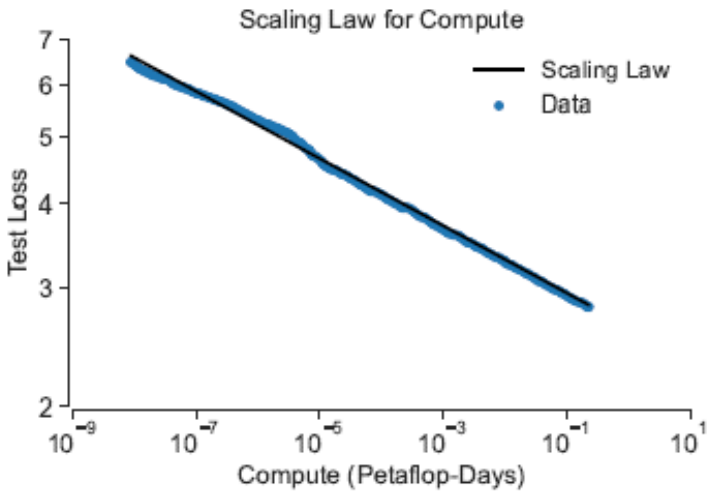


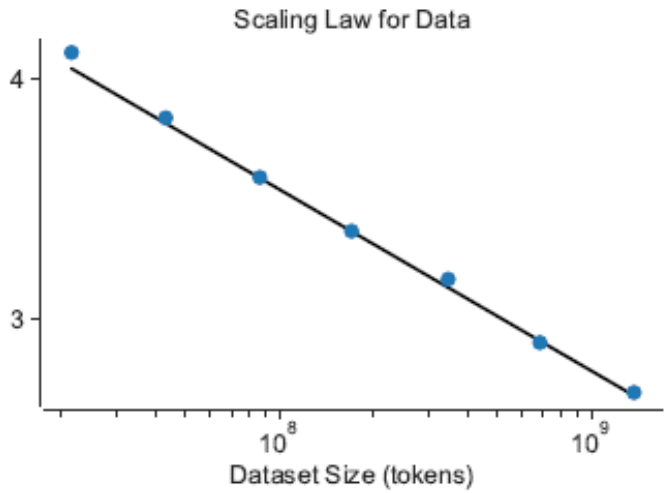


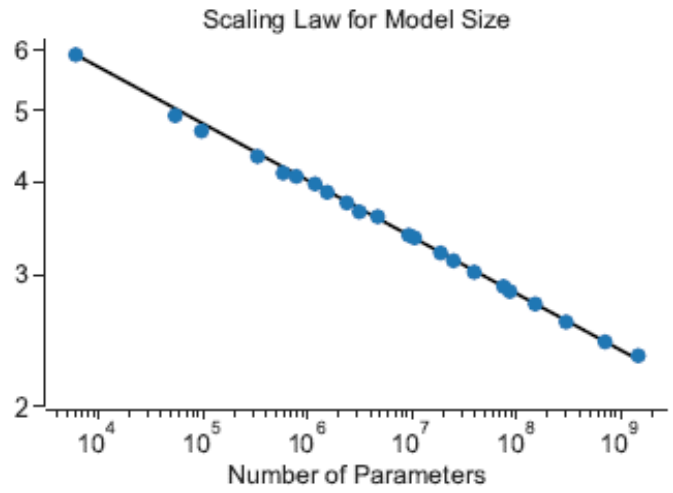


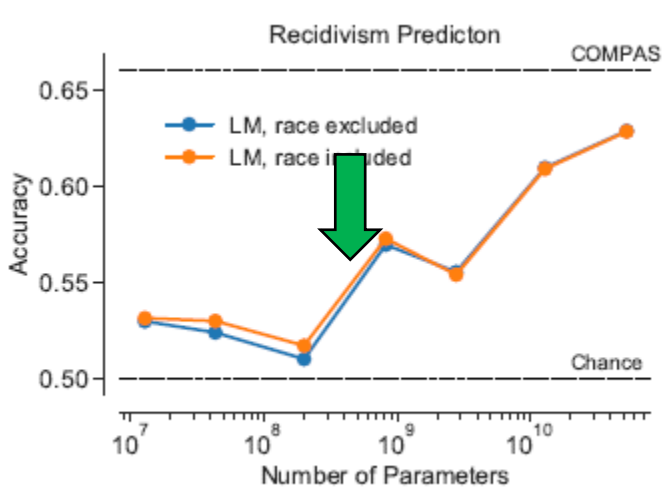


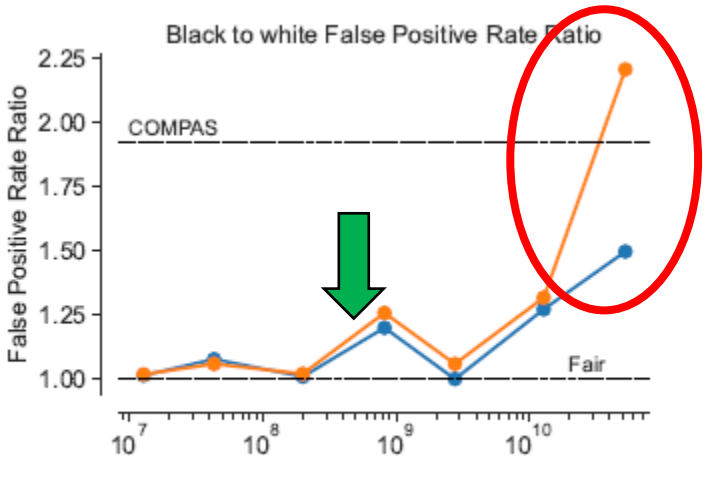
That Goes for Failures Too

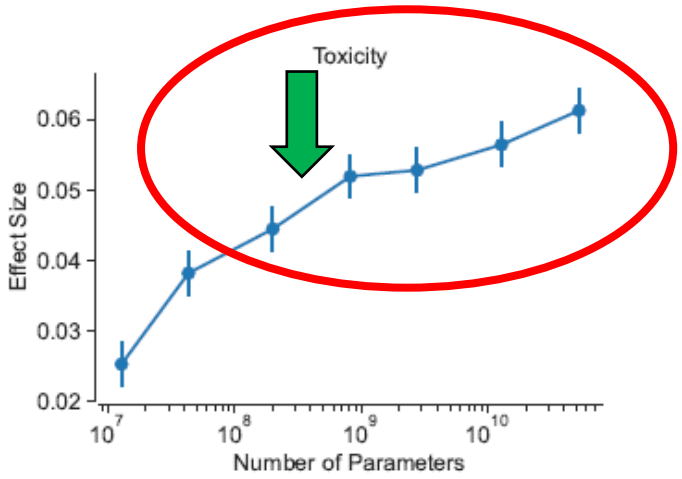




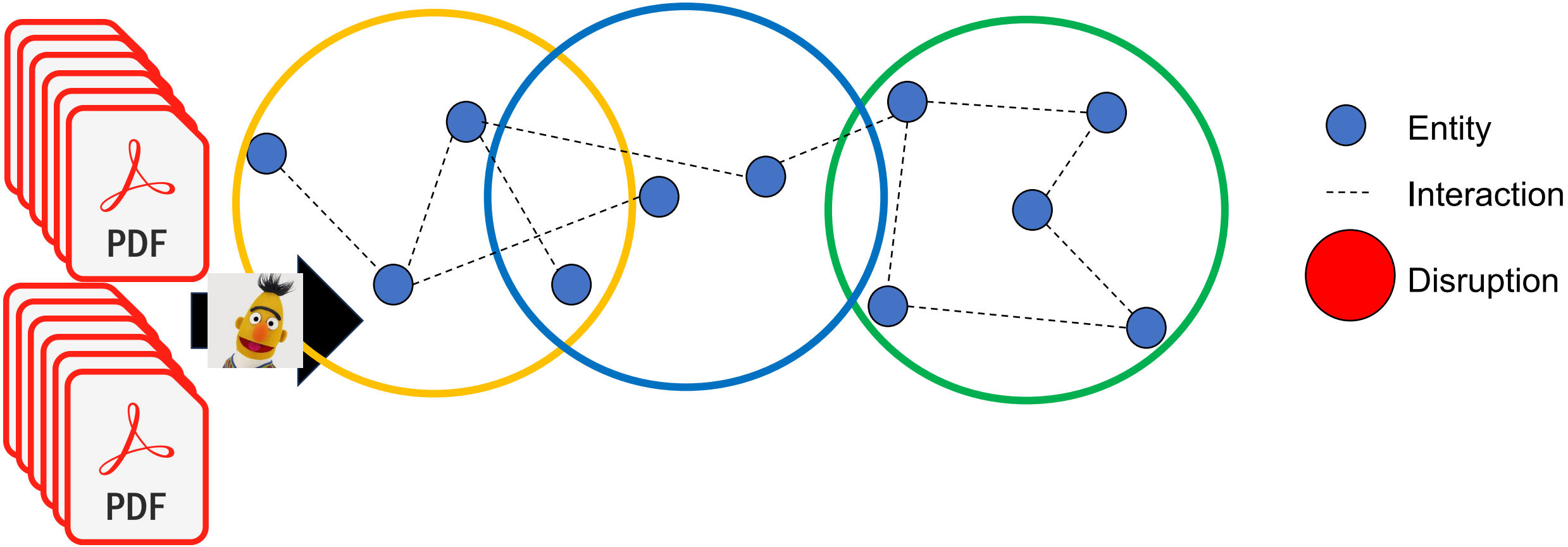








Sanity check:
Ability to detect topic with extracted entities



Methods (models)

- 4 major types
- Document segmented to fully fit the attention windows
- At most 100 entities extracted
  - Select by highest confidence

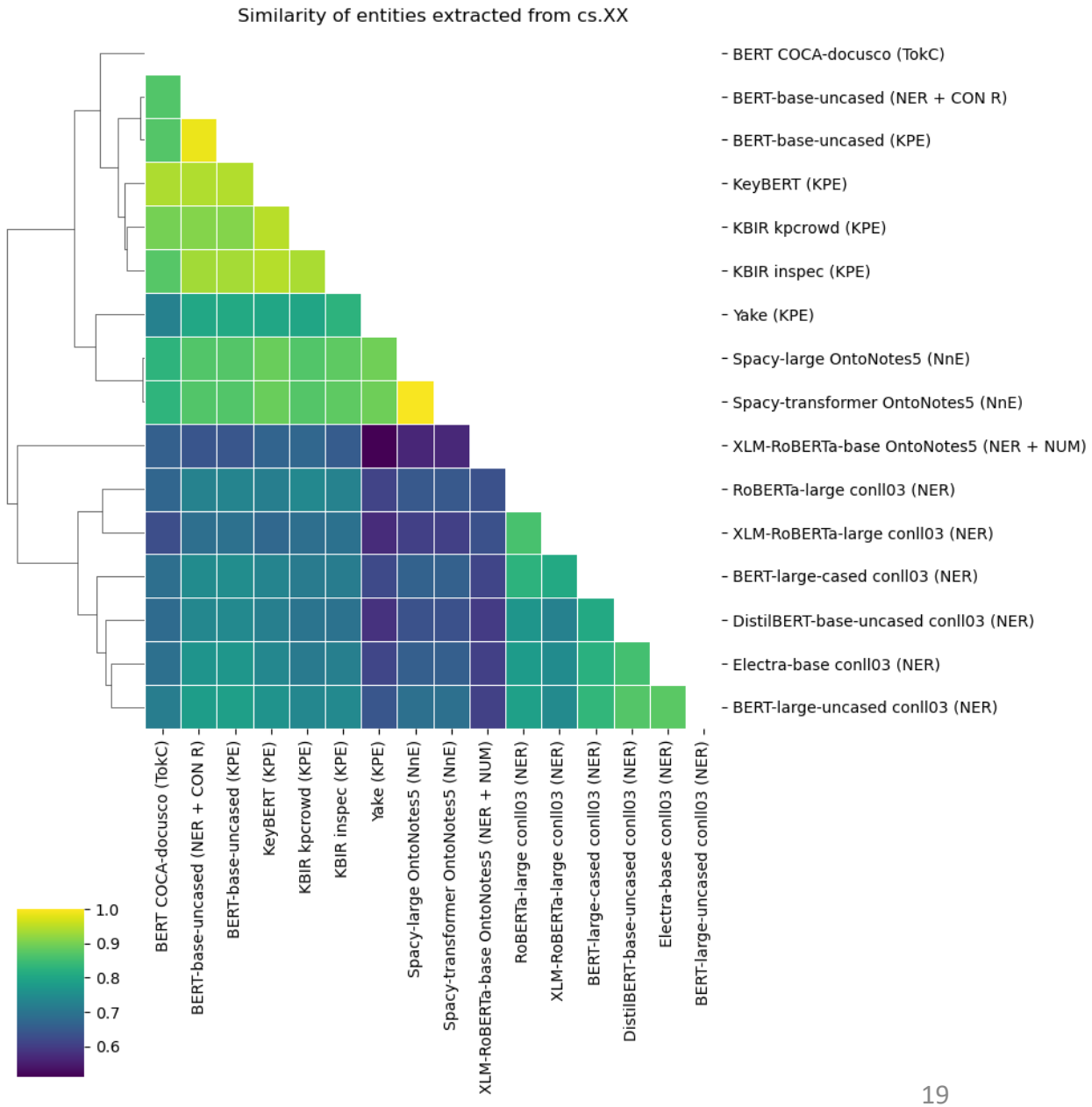
Model Name	Refs	Entities/Doc	Type
spaCy Large*^{p}	[17]	99.3 ± 6.93	Noun Extractor
spaCy Transformer^p	[17]	99.3 ± 6.97	
Yake*^{p}	[5]	19.9 ± 1.97	Keyphrase Extractor
KeyBERT^p	[15]	99.3 ± 7.25	
KBIR kpcrowd	[23, 25]	96.9 ± 14.6	
KBIR inspec	[23, 37]	76.4 ± 27.7	
BERT-base-uncased	[11]	44.7 ± 24.0	
BERT-base-uncased	[11]	43.3 ± 23.3	NER+CON R
XLM-RoBERTa-base Onconotes 5	[40, 18]	36.4 ± 23.4	NER+NUM
ELECTRA-base conll03	[8, 38]	39.9 ± 25.0	NER
BERT-large-cased conll03	[11, 38]	41.7 ± 24.9	
BERT-large-uncased conll03	[11, 38]	33.5 ± 23.3	
DistilBERT-base-uncased conll03	[39, 38]	37.7 ± 24.8	
RoBERTa-large conll03	[24, 38]	28.7 ± 21.1	
XLM-RoBERTa-large conll03	[14, 38]	26.0 ± 19.5	
BERT COCA-docusco	[11, 20]	99.6 ± 6.11	TokC

Methods (visualisations)

- Hierarchical clustering
  - Embedded with SpaCy
  - Average cosine distance
  - Identify similitude between extractor
- 2D Projection
  - Subsample data: reduce processing time and number of point
  - 6 embeddings:
    - SpaCy, GloVe, Fasttext, Word2Vec, BERT-large, GPT-2
  - 4 low-dimensional projection
    - Linear, spectral, t-SNE, UMAP
  - Show if themes can be detected in an unsupervised way

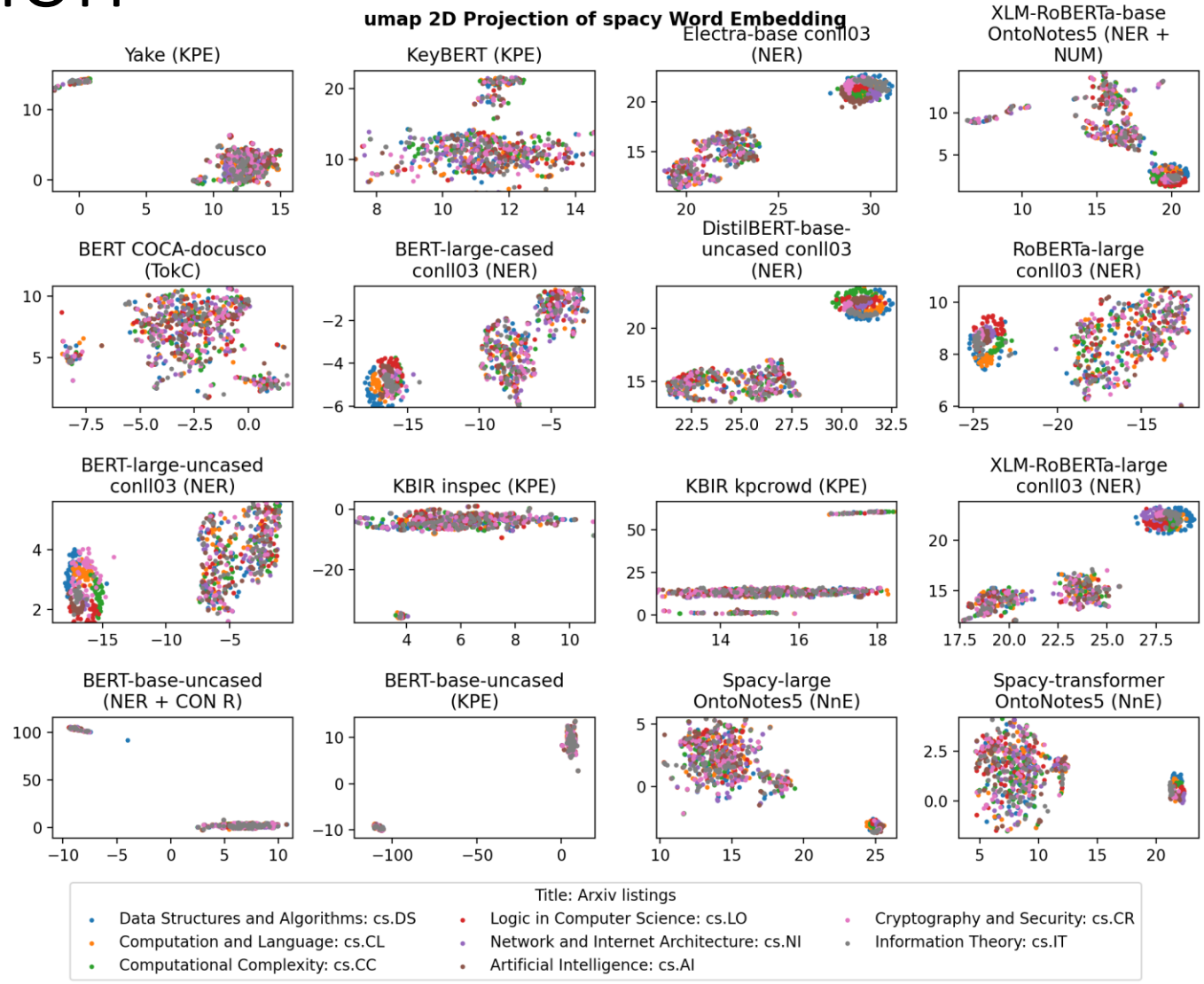
Results and Discussion

- Performance manly define by architecture and fine-tuned dataset
- Dataset not based on scientific texts
  - Conll03
- => Not suited for scientific articles

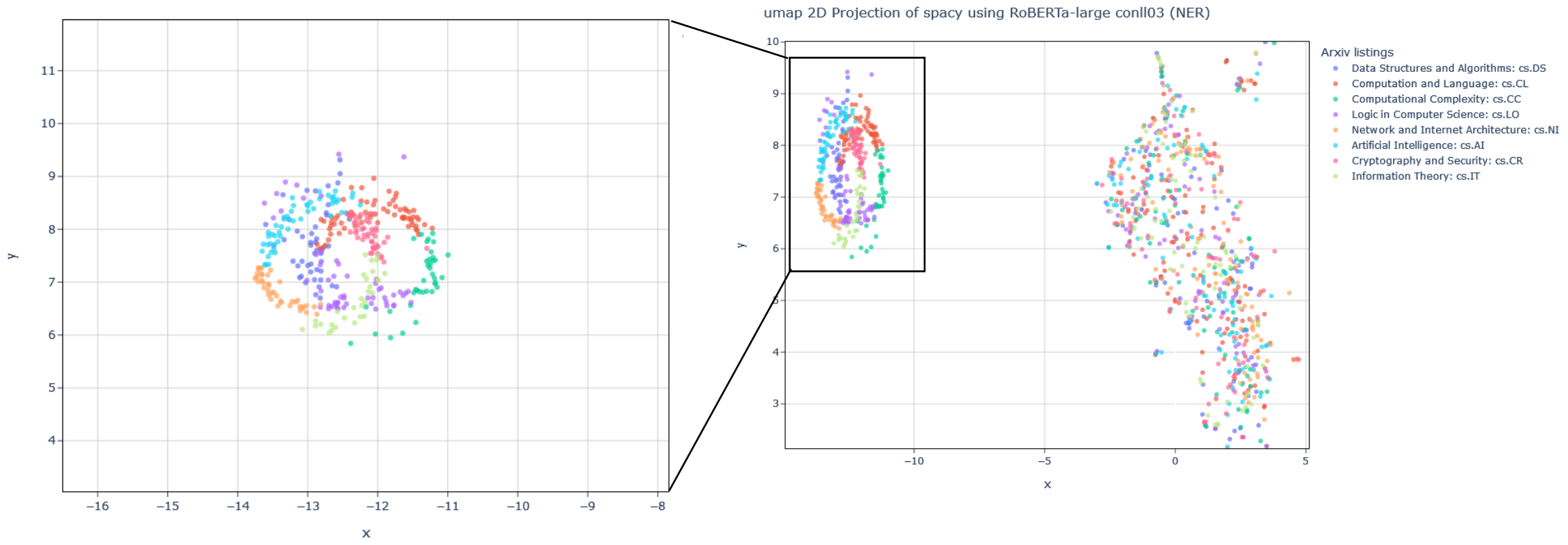


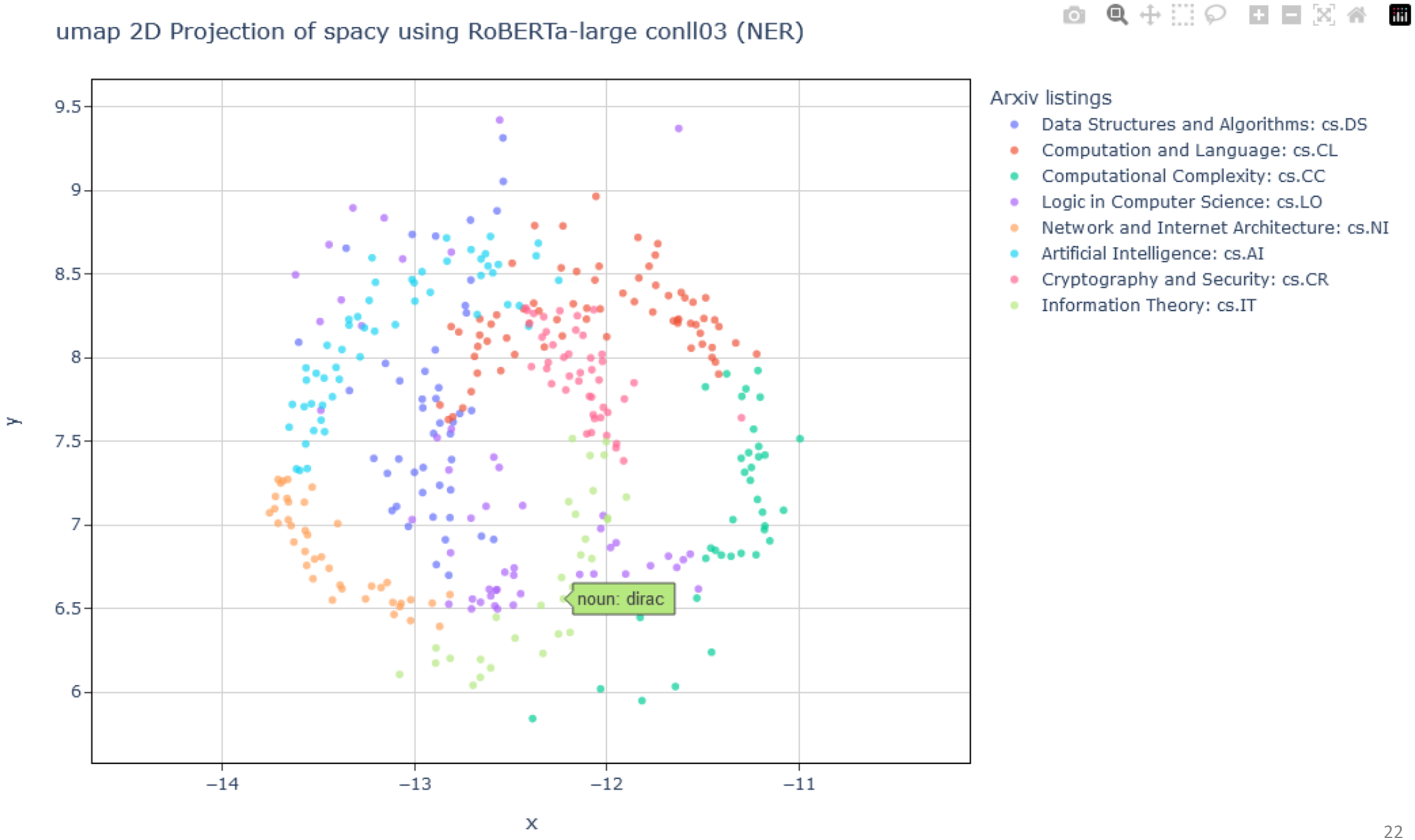
Results and Discussion

- Cosine similarity of embedding do not perform well to cluster themes
  - Even with 2D embedding algorithm that tend to overfit
- Exception with NER



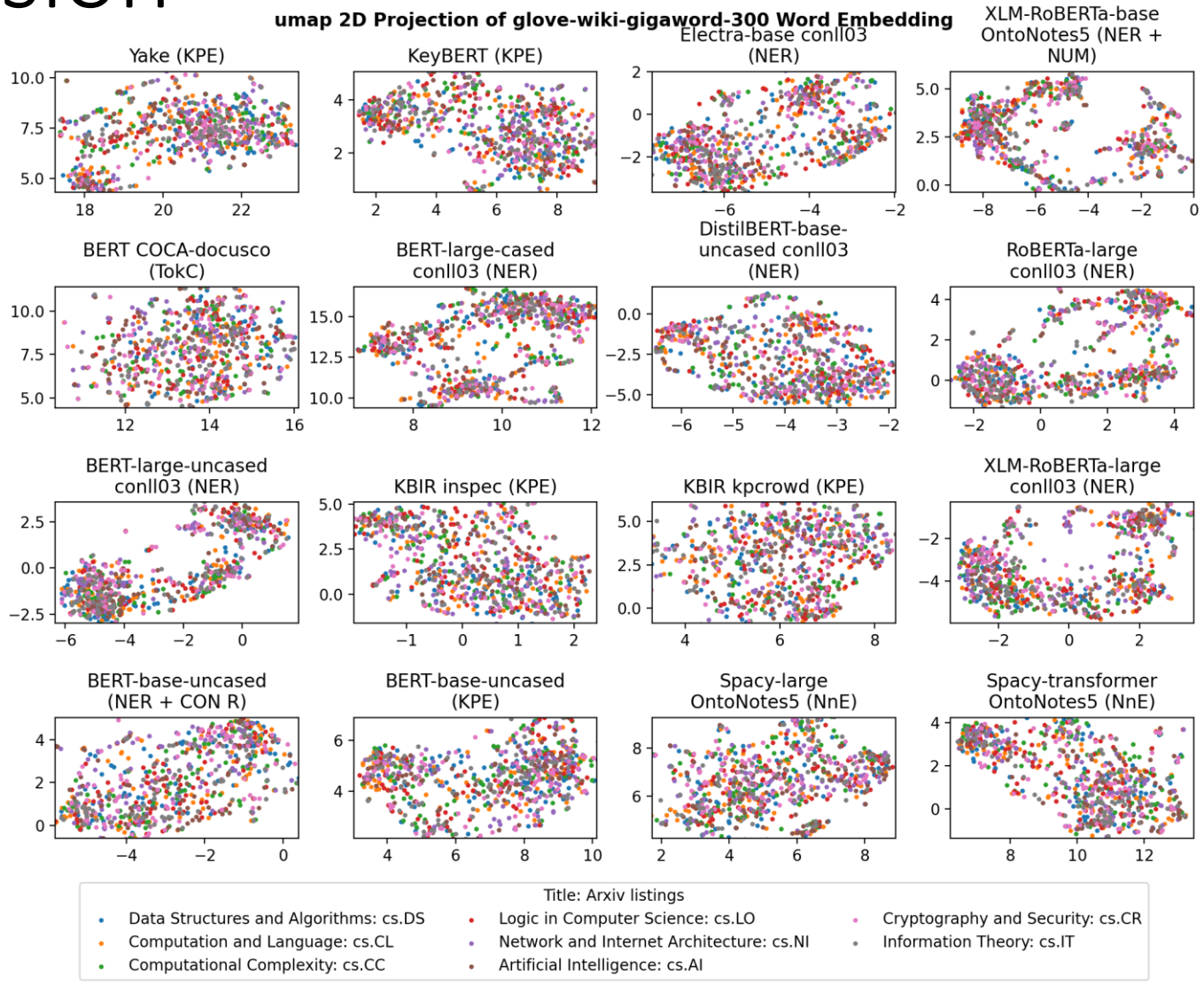
Umap projection of Spacy using RoBERTa-large conll03 (NER)



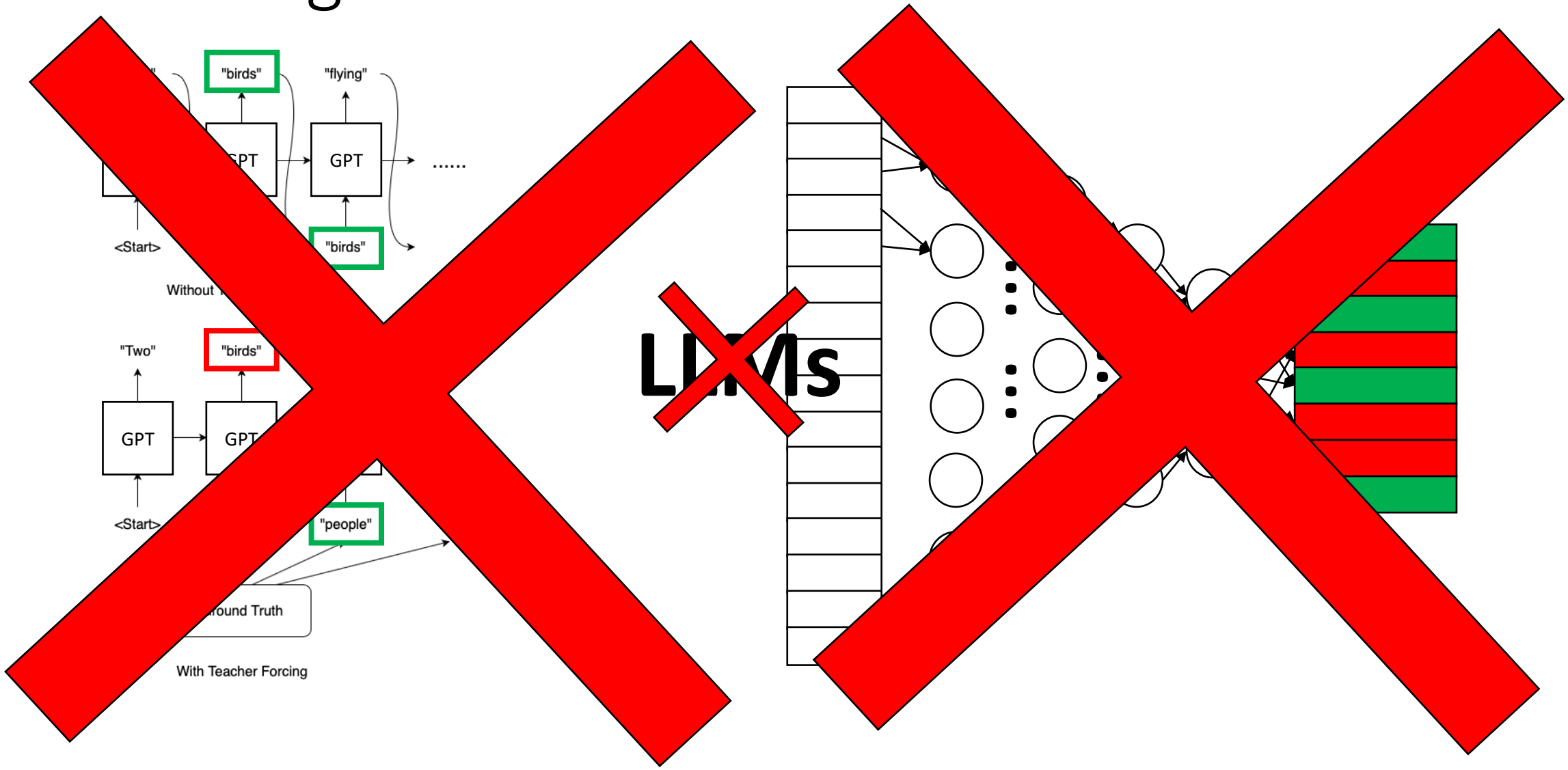

umap 2D Projection of spacy using RoBERTa-large conll03 (NER)
Arxiv listings
Data Structures and Algorithms: cs.DS
Computation and Language: cs.CL
Computational Complexity: cs.CC
Logic in Computer Science: cs.LO
Network and Internet Architecture: cs.NI
Artificial Intelligence: cs.AI
Cryptography and Security: cs.CR
Information Theory: cs.IT
noun: dirac
9.5
9
8.5
8
7.5
7
6.5
6
y
−14
−13
−12
−11
x

Results and Discussion

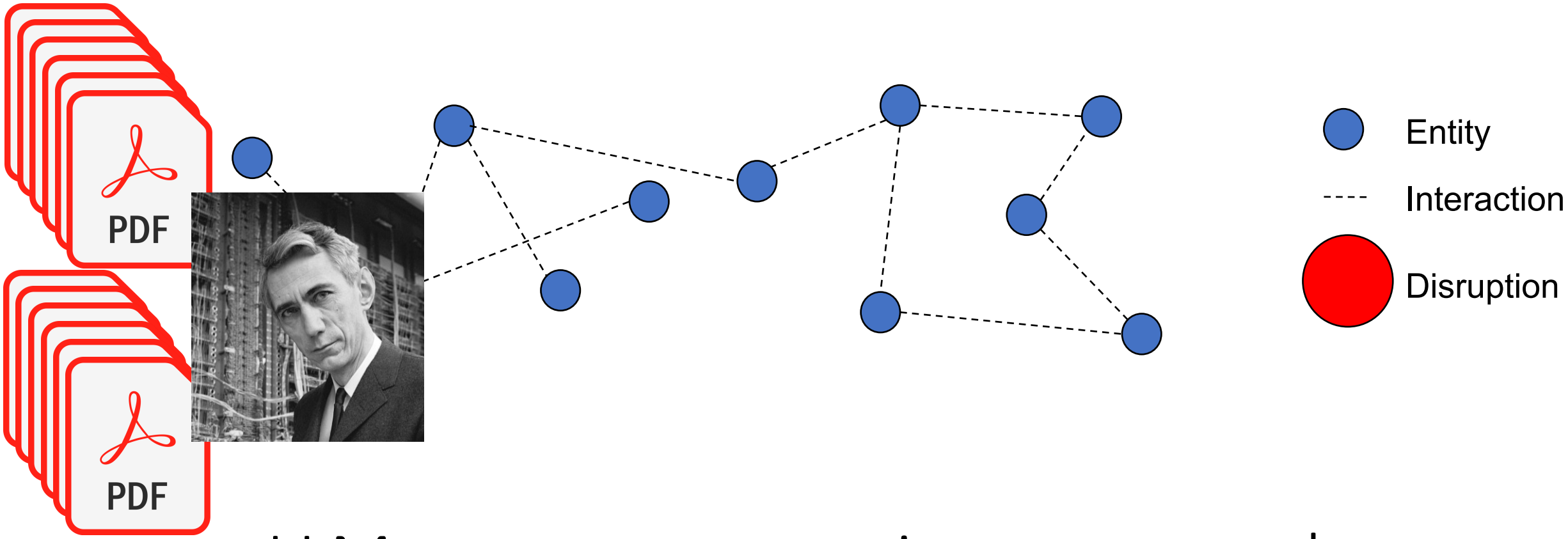
- Cosine similarity highly dependent of embedding space
- Important change with different embedding and algorithm



Where go from there?



WP3: Enabling rigorous indirect technological impact prediction

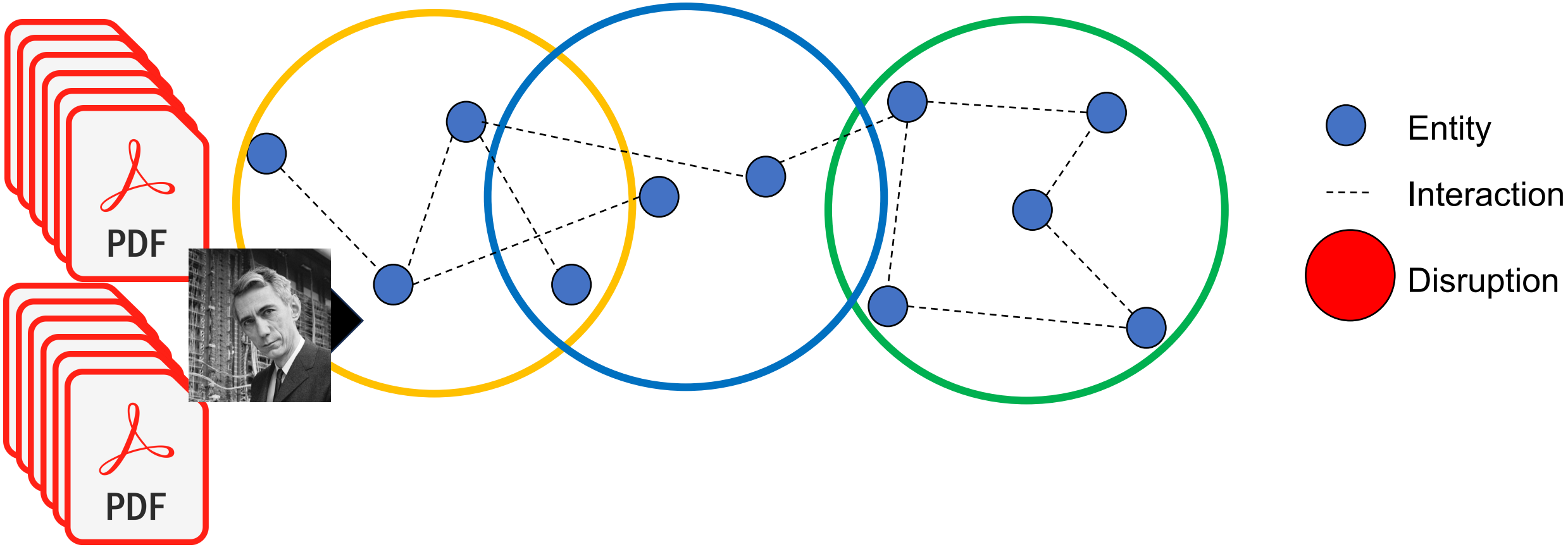


LLMs are ~ compressive autoencoders

Let’s build compressive autoencoders directly on what we are interested in!

1. Concepts:
   - Extract **all** nouns
2. Specific to cyber-security:
   - Find **specific** nouns
3. That are connected:
   - Find **correlated** nouns

Sanity check:
Ability to detect topic with extracted entities



Extract nouns

Approach

- PDF:
  - Filter for English
  - Remove header and bibliography
  - Transform to text
- spaCy
  - Pull noun groups
  - (high + school vs high school)
- Retain:
  - Nouns found in >4 documents
  - Nouns found >4 times

Sample:

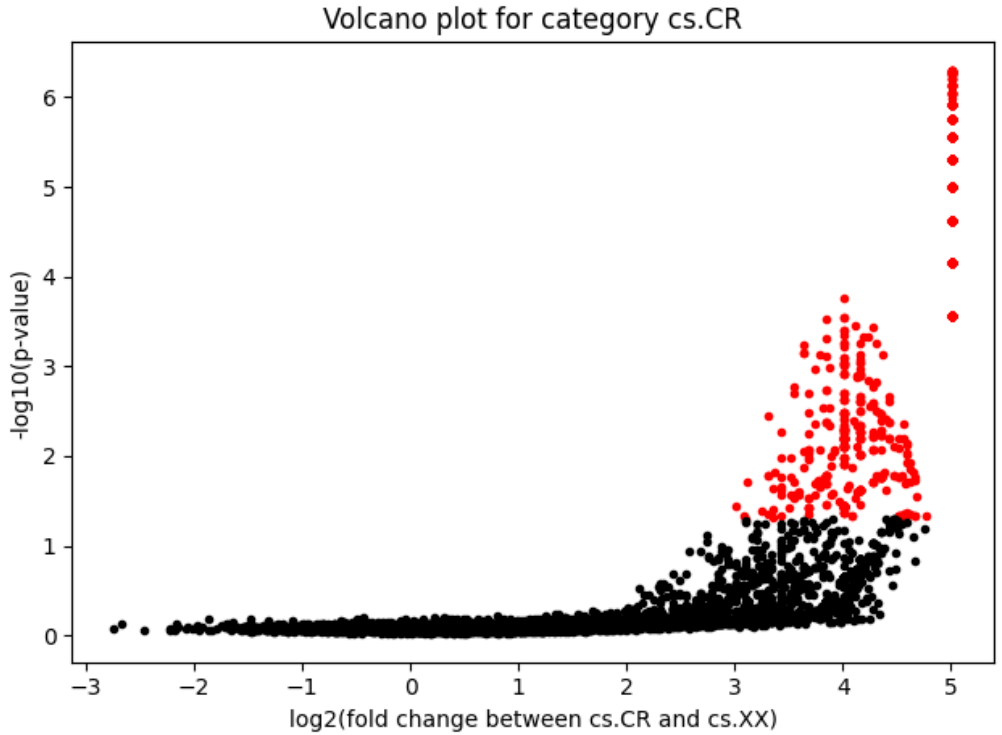
- secret sharing scheme
- pseudorandom function
- residual codekey agreement
- pseudorandom permutation
- geometric code
- non-negligible advantage
- gf q rational divisor
- knowledge property
- new cryptosystem

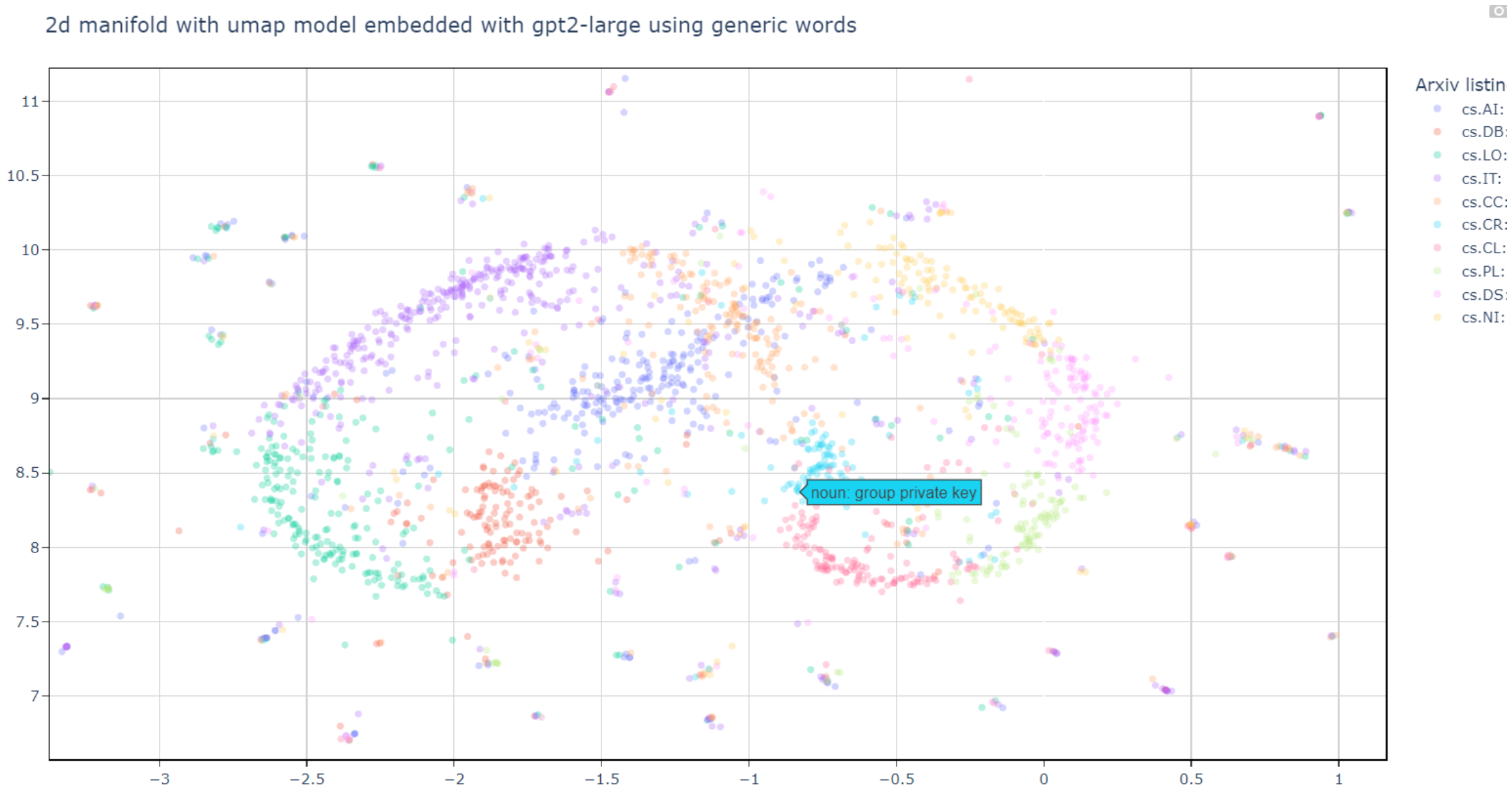
Find specific nous

Approach

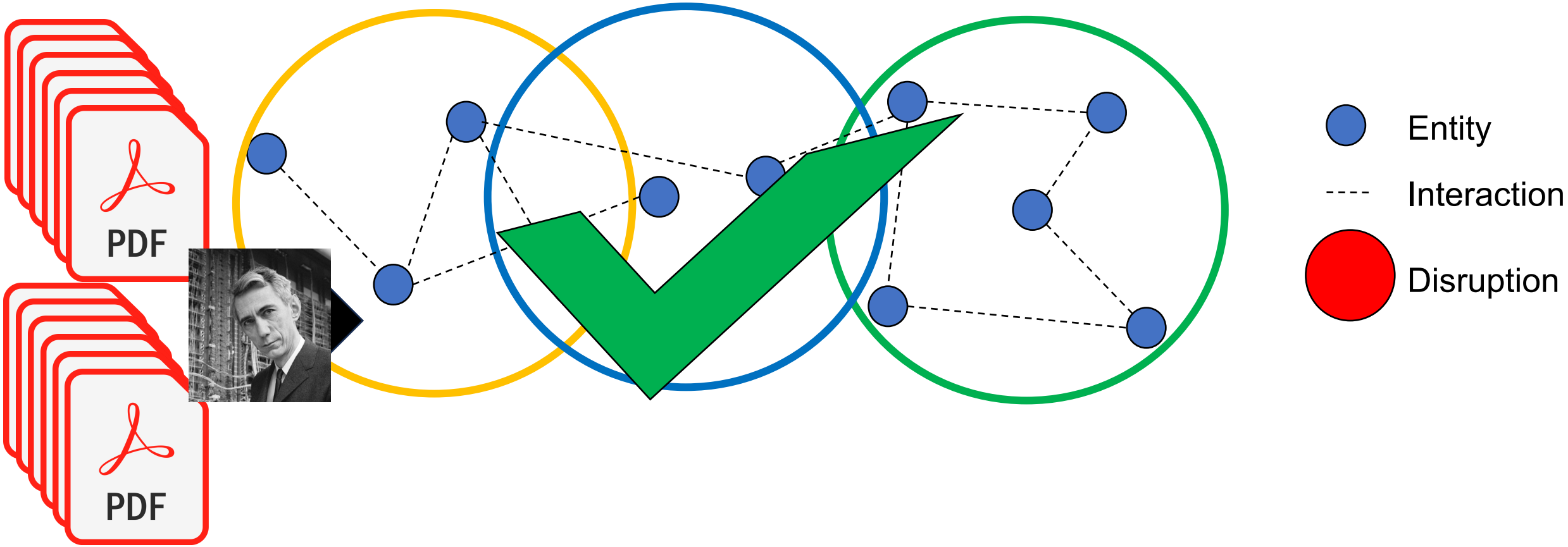
- Nouns frequency in:
  - arXiv cs.CR
  - arXiv cs
  - BookCorpus
- Compare frequencies
  - Classical Statistics

Results

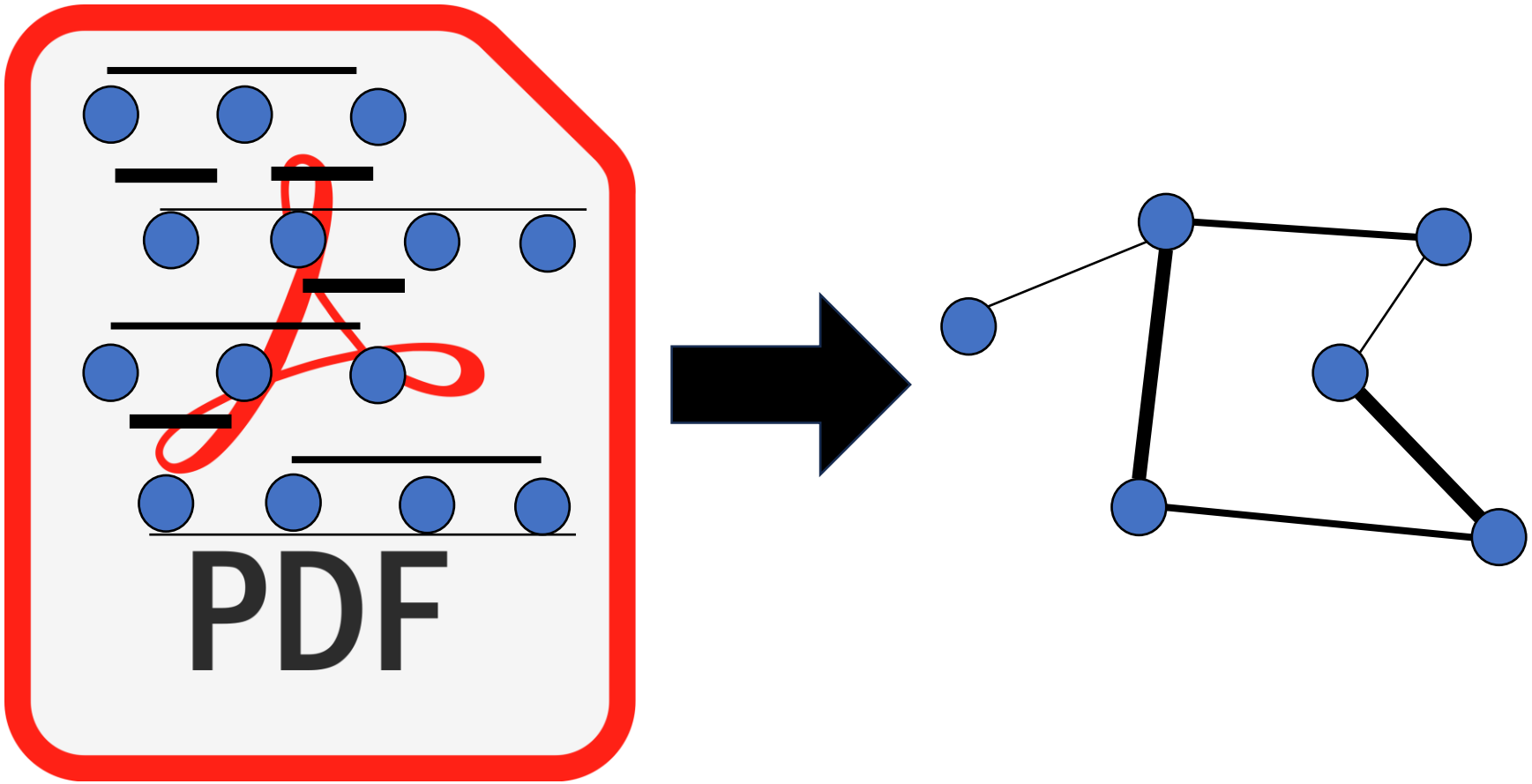



2d manifold with umap model embedded with gpt2-large using generic words
noun: group private key
Arxiv listin
cs.AI:
cs.DB:
cs.LO:
cs.IT:
cs.CC:
cs.CR:
cs.CL:
cs.PL:
cs.DS:
cs.NI:
11
10.5
10
9.5
9
8.5
8
7.5
7
−3
−2.5
−2
−1.5
−1
−0.5
0
0.5
1

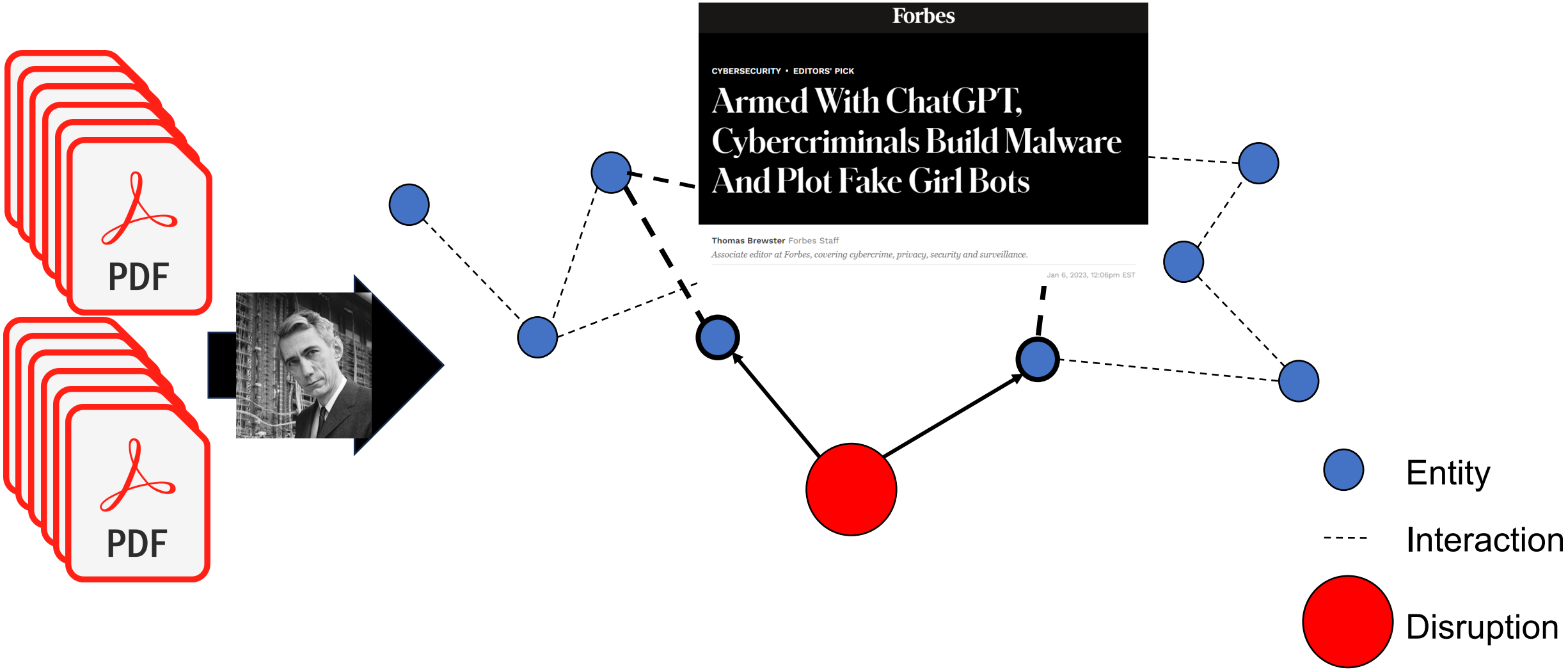
Sanity check:
Ability to detect topic with extracted entities



Next stop: Term relations



Instead of a conclusion



CYD | CYBER DEFENCE CAMPUS

Maxime **Würsch**
CYD intern

Dimitri **Percia David**
Assistant professor UAS

Andrei **Kucharavy**
Research associate UAS

EPFL

Hes·so VALAIS WALLIS

Thanks for your attention!

Questions?

Gen Learning Center:
https://tinyurl.com/hevs-gen-learning
Report: https://arxiv.org/abs/2303.12132