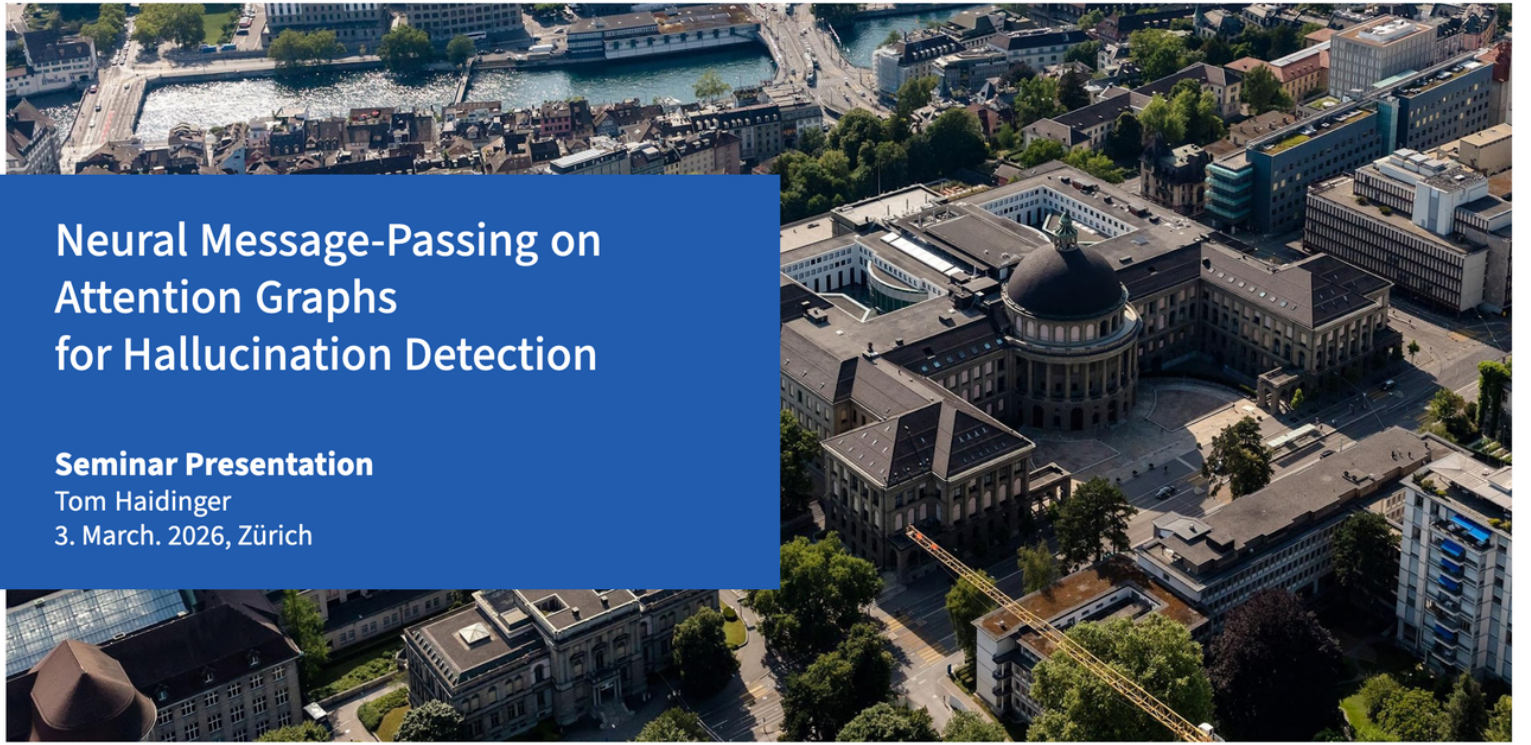
ETH zürich



Neural Message-Passing on Attention Graphs for Hallucination Detection

Seminar Presentation

Tom Haidinger

3. March. 2026, Zürich

Supermarket AI meal planner app suggests recipe that would create chlorine gas – The Guardian

Lawyers blame ChatGPT for tricking them into citing bogus case law

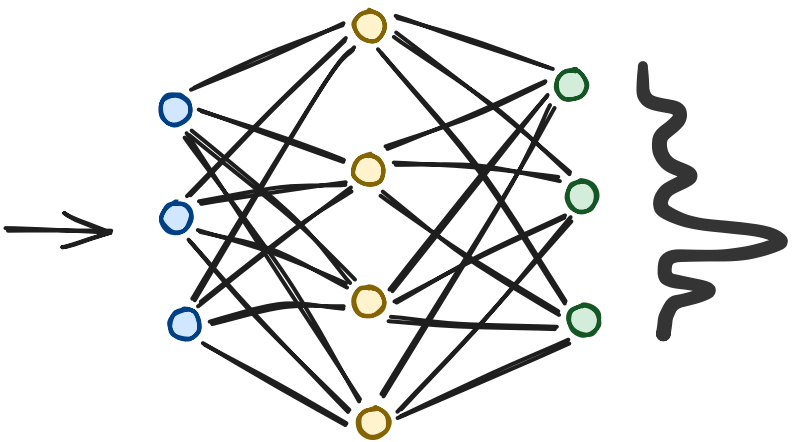
– AP News

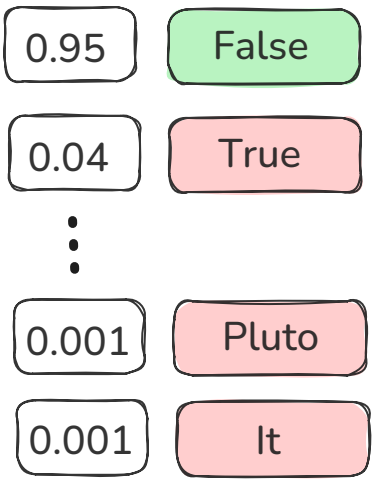
NEDA Suspends AI Chatbot for Giving Harmful Eating Disorder Advice

– Psychiatrist.com

Incorrect Answer in Distribution

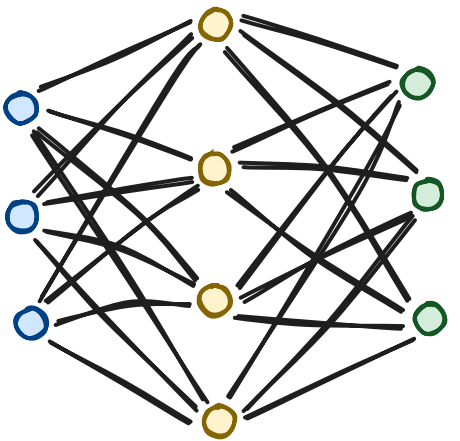
True or False, is Pluto a planet?





Incorrect Mode > Correct Mass

Pluto ...

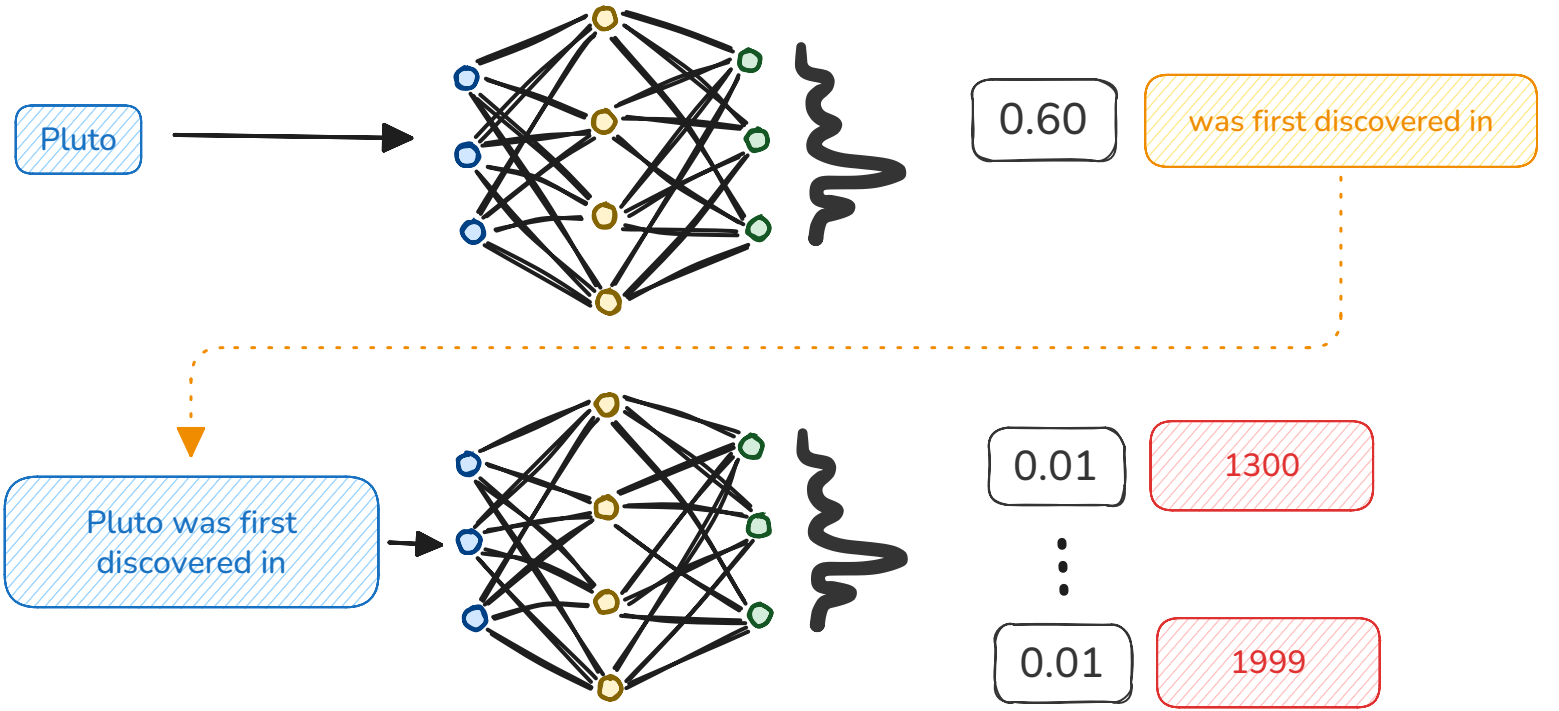


0.10 is the smallest planet in the Solar System

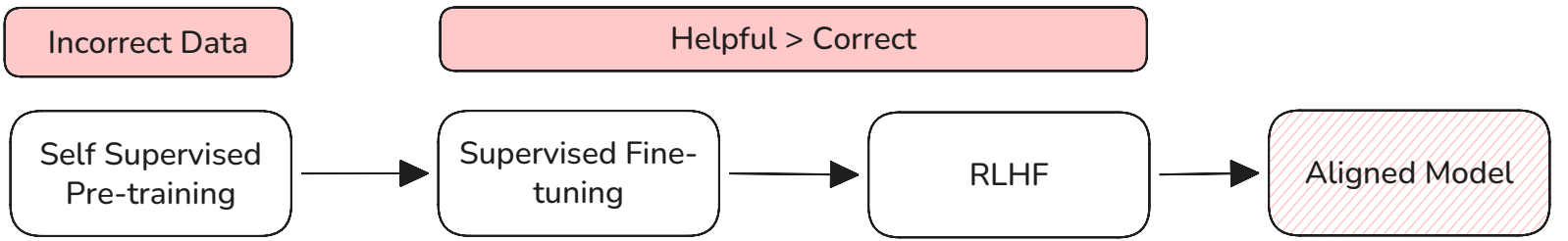
0.08 has an icy surface

0.08 was reclassified as a dwarf planet in 2006

Correct path can lead to incorrect outcome



Dataset & Training



Contextual Hallucination

Prompt Context

Schedule

Date	Presenter	Title	Mentor	Slides
February 17	Frédéric Berdoz	Introduction to Scientific Presentations	-	[pdf]
February 24	Jonathan Maillefaud	Generative Graph Pattern Machine	Saku Peltonen	[pdf]
	-	TBA	-	-
March 03	Tom Haidinger	Neural Message-Passing on Attention Graphs for Hallucination Detection	Florian Grötschla	-
	Yucheng Wang	MiMo-Audio: Audio Language Models are Few-Shot Learners	Luca Lanzendörfer	-

LLM: Florian is presenting today!

Non-Contextual Hallucination

Training Data

Seminar in Deep Neural Networks (FS 2025)

Organization

When & Where: Tuesdays 10:15 - 12:00, ETZ E 9.
First seminar: 18.02.2025
Last seminar: 27.05.2025
Easter holidays: No seminar on 22.04.2025
Coordinators: Frédéric Berdoz, Benjamin Estermann, Roger Wattenhofer.

2025

Prompt Context

2026

Title
Introduction to Scientific Presentations
Generative Graph Pattern Machine
TBA
Neural Message-Passing on Attention Graphs for Hallucination Detection
MiMo-Audio: Audio Language Models are Few-Shot Learners

LLM: Prof. Wattenhofer is presenting today!

Contextual

Given: CNN + Daily Mail Articles

Question: Jeremy Clarkson, has been X as presenter on Top Gear

fired	(0.0)
canceled	(0.1)
celebrated	(0.9)
promoted	(1.0)

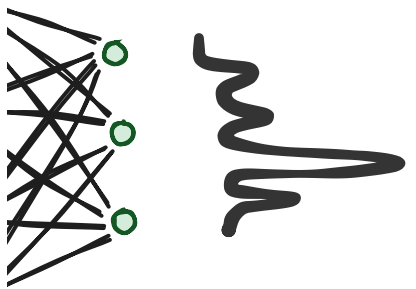
Non-Contextual

Movies Dataset

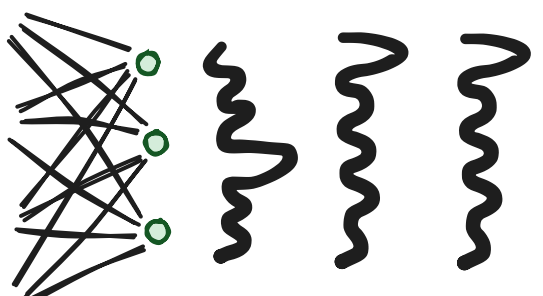
What actor played Han Solo in Star Wars?

Harrison Ford	(0.0)
Harrison Page	(0.5)
George Lucas	(0.9)

Grey Box Approach



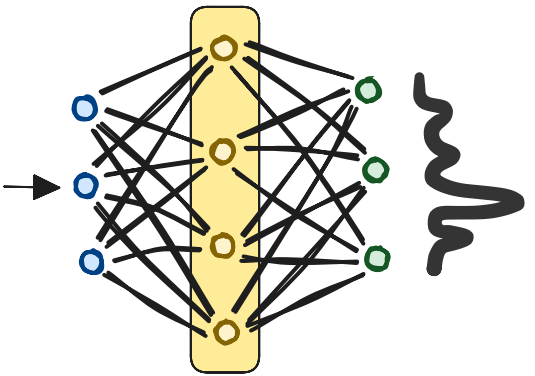
Entropy/ uncertainty metrics

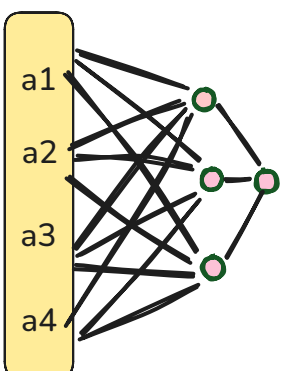


Multiple samples per prompt

White Box Approach – Activation Probes

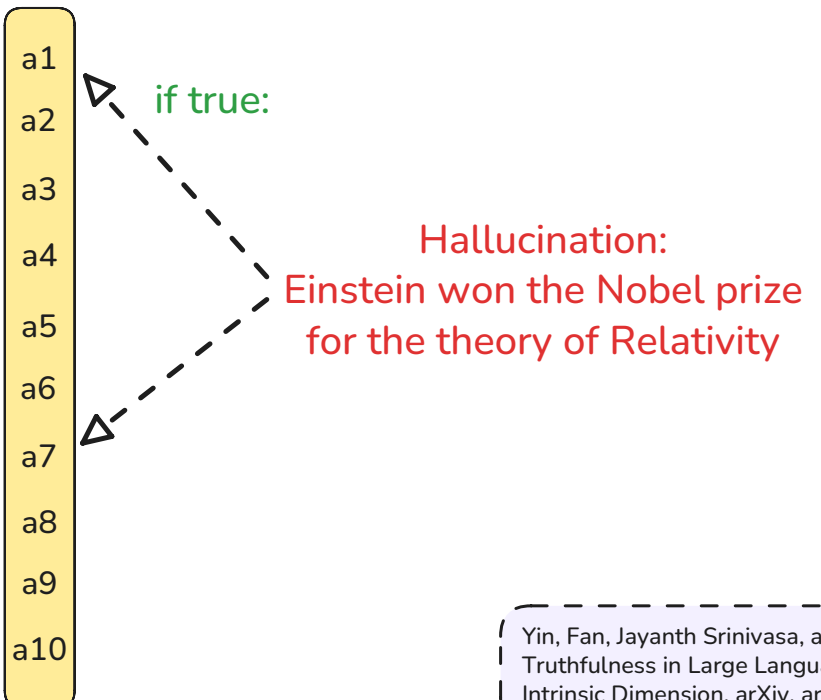
take activations of last token





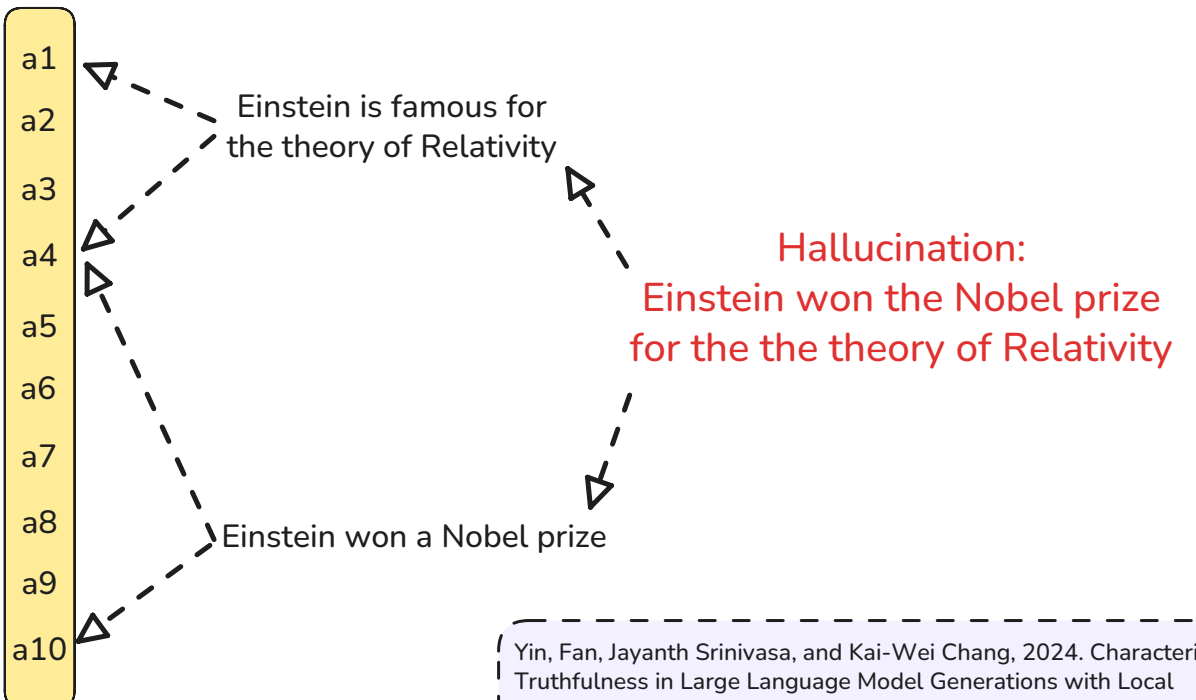
G. Alain and Y. Bengio, 2017. “Understanding intermediate layers using linear classifier probes,” arXiv preprint arXiv:1610.01644.

Intrinsic Dimention



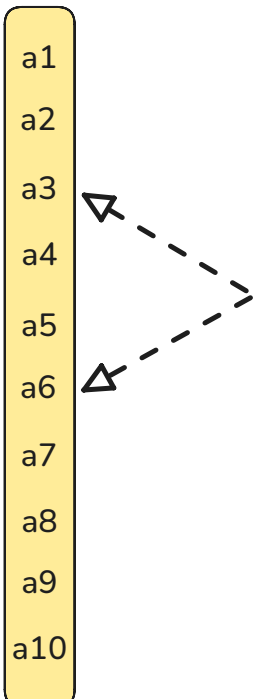
Yin, Fan, Jayanth Srinivasa, and Kai-Wei Chang, 2024. Characterizing Truthfulness in Large Language Model Generations with Local Intrinsic Dimension. arXiv, arXiv:2402.18048.

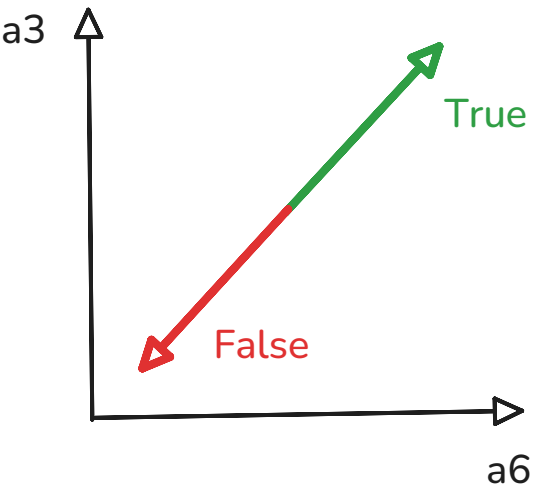
Intrinsic Dimention



Yin, Fan, Jayanth Srinivasa, and Kai-Wei Chang, 2024. Characterizing Truthfulness in Large Language Model Generations with Local Intrinsic Dimension. arXiv, arXiv:2402.18048.

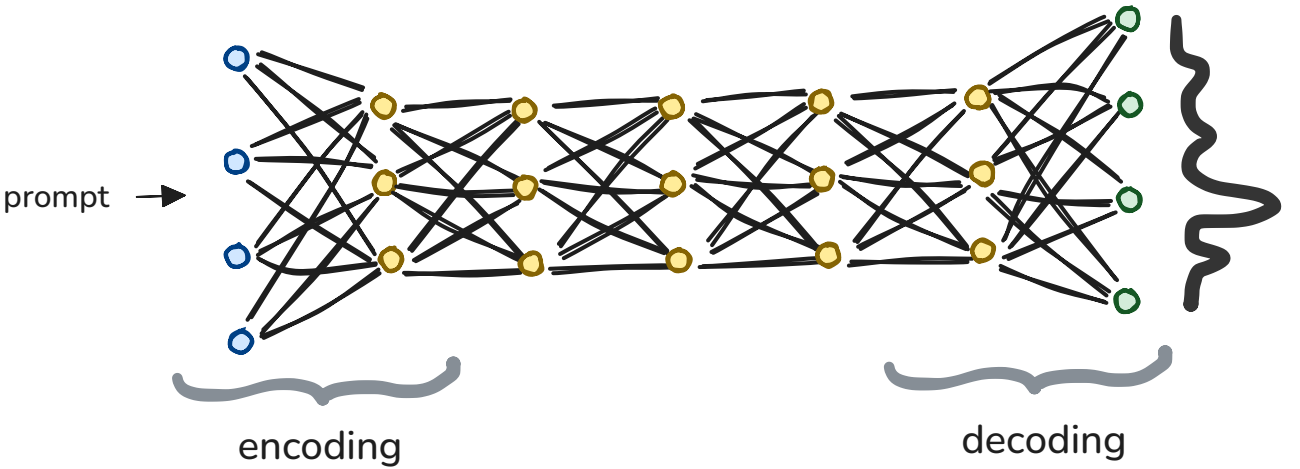
'Truth' Vectors



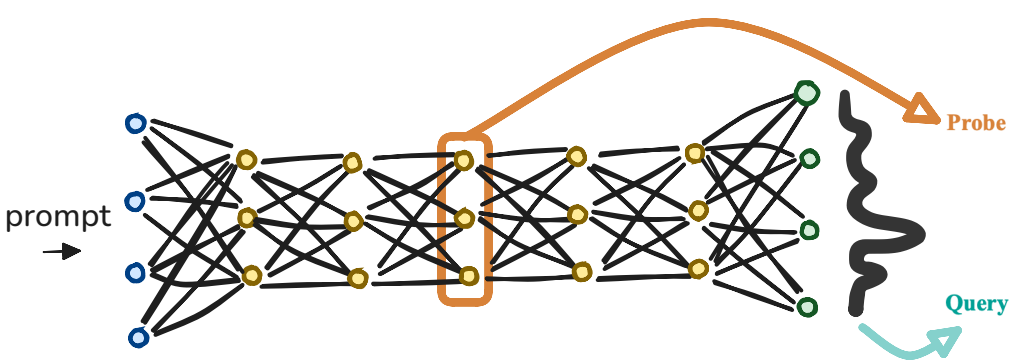


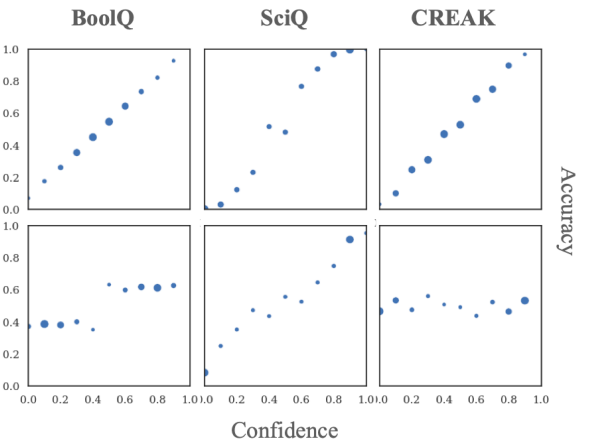
Azaria, Amos, and Tom Mitchell, 2023. The Internal State of an LLM Knows When It's Lying. Association for Computational Linguistics, 2023, pp. 967–976

Optimal Probing Layer



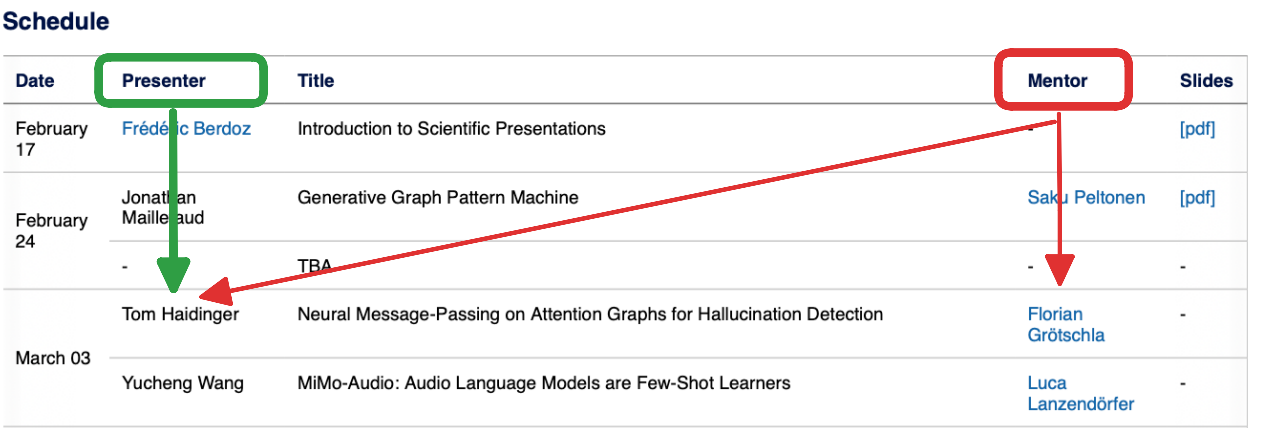
Optimal Probing Layer





Liu, Kevin, et al. 2023. Cognitive Dissonance,: Why Do Language Model Outputs Disagree with Internal Representations of Truthfulness? arXiv, arXiv:2312.03729.

Attention Intuition Ex. 1



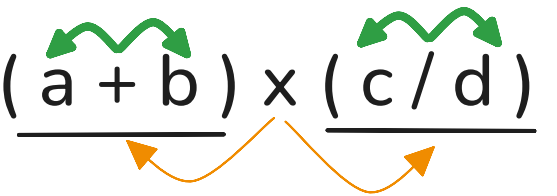
Schedule

Date	Presenter	Title	Mentor	Slides
February 17	Frédéric Berdoz	Introduction to Scientific Presentations	-	[pdf]
February 24	Jonathan Mailleaud	Generative Graph Pattern Machine	Saku Peltonen	[pdf]
	-	TBA	-	-
March 03	Tom Haidinger	Neural Message-Passing on Attention Graphs for Hallucination Detection	Florian Grötschla	-
	Yucheng Wang	MiMo-Audio: Audio Language Models are Few-Shot Learners	Luca Lanzendörfer	-

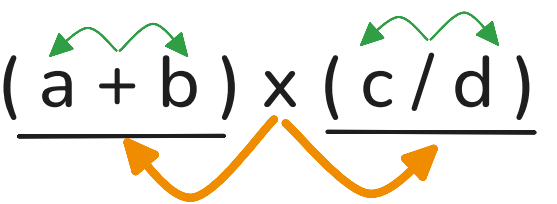
Encoding: " .... < Presenter | Title | Mentor > ..... < Tom | Paper Title | Florian >"

Attention Intuition Ex. 2

Step 1:



Step 2:



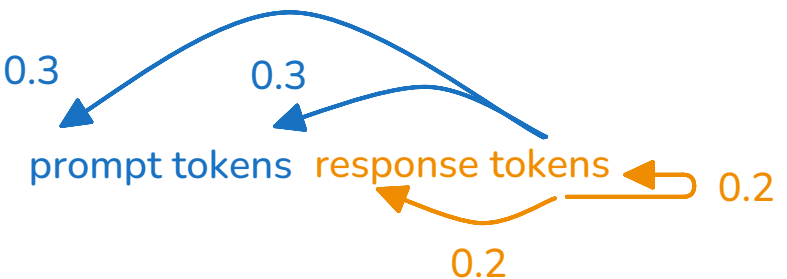
Attention Intuition Ex. 3

```
def foo(x, y):
x = doo(x)
y = doo(y)
return x+y
```

```
def doo(z):
return z.size()
```

LookbackLens

$$\ell_i^{l,h} = \frac{P_i^{l,h}}{P_i^{l,h} + R_i^{l,h}}.$$



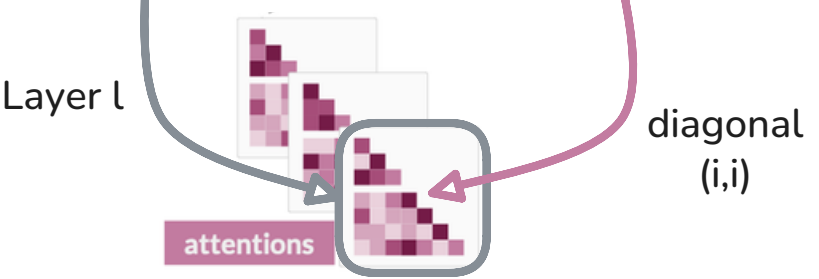
$$l_4^l, h = \frac{0.3 + 0.3}{0.3 + 0.3 + 0.2 + 0.2} = 0.6$$

Chuang, Yung-Sung, et al, 2024. Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using Only Attention Maps. arXiv, arXiv:2407.07071.

Non-Contextual Response Level
LLM-Check

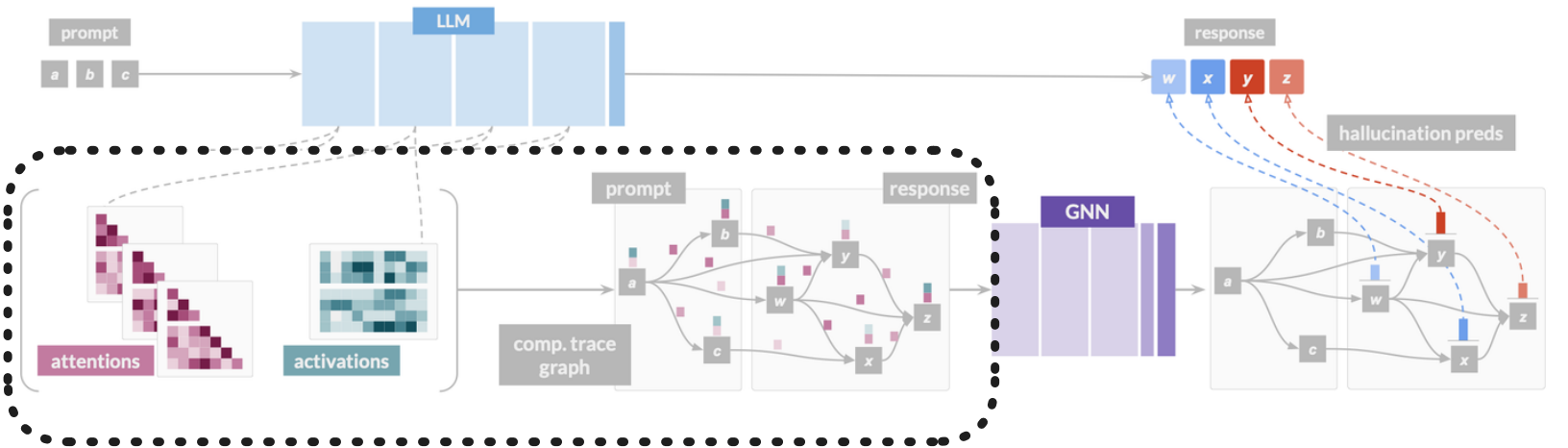
Attention Score at layer l:
for token i

$$c^l = \frac{1}{H} \sum_{h=0}^{H-1} \sum_{i=0}^{n-1} \log(\alpha_{i,i}^{l,h}),$$



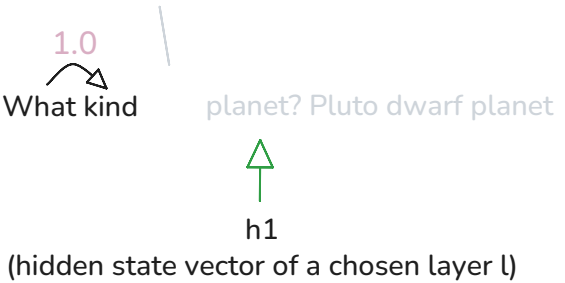
Sriramanan, Gaurang, et al, 2024. “LLM-Check: Investigating Detection of Hallucinations in Large Language Models.” The Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS 2024),

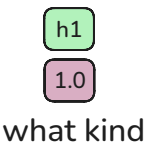
CHARM: Catching HAllucinated Responses via (learnable) Message-passing



Example

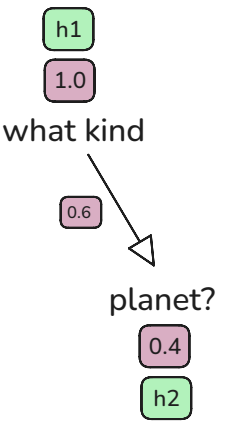
What kind of planet is Pluto? –> A dwarf planet





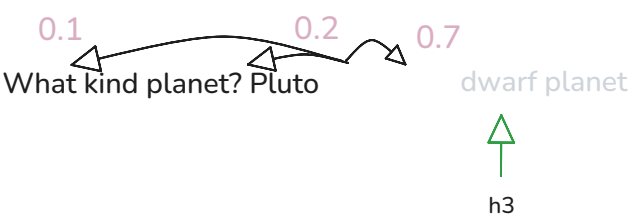
Example

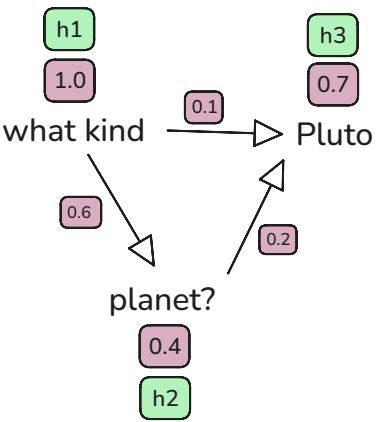
What kind of planet is Pluto? –> A dwarf planet



Example

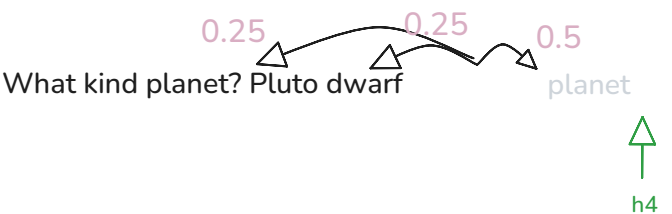
What kind of planet is Pluto? –> A dwarf planet

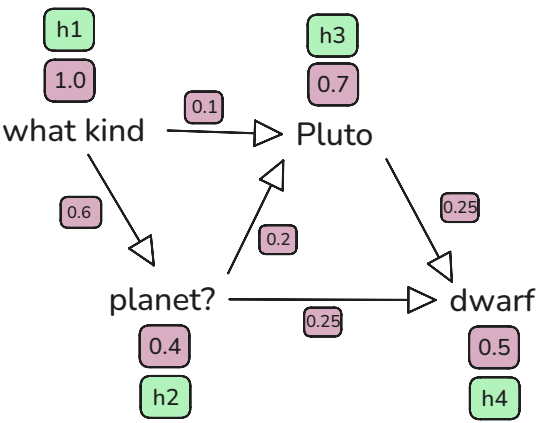




Example

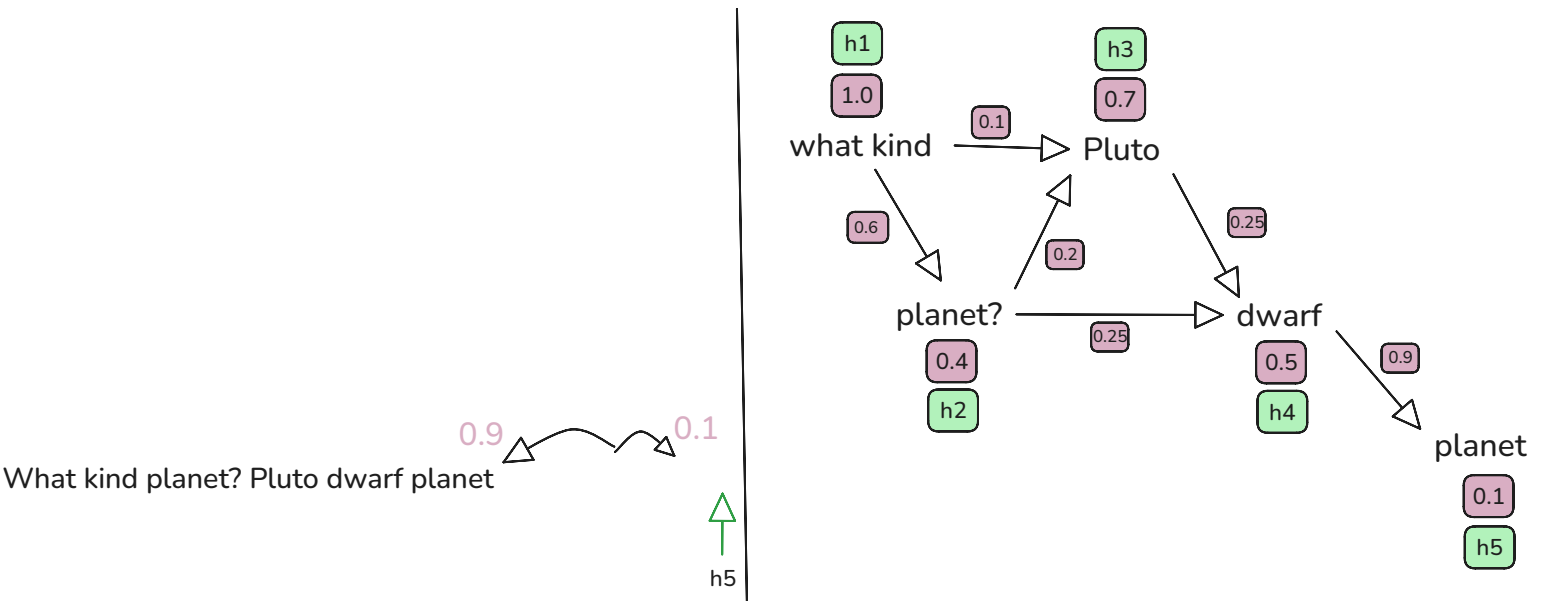
What kind of planet is Pluto? –> A dwarf planet



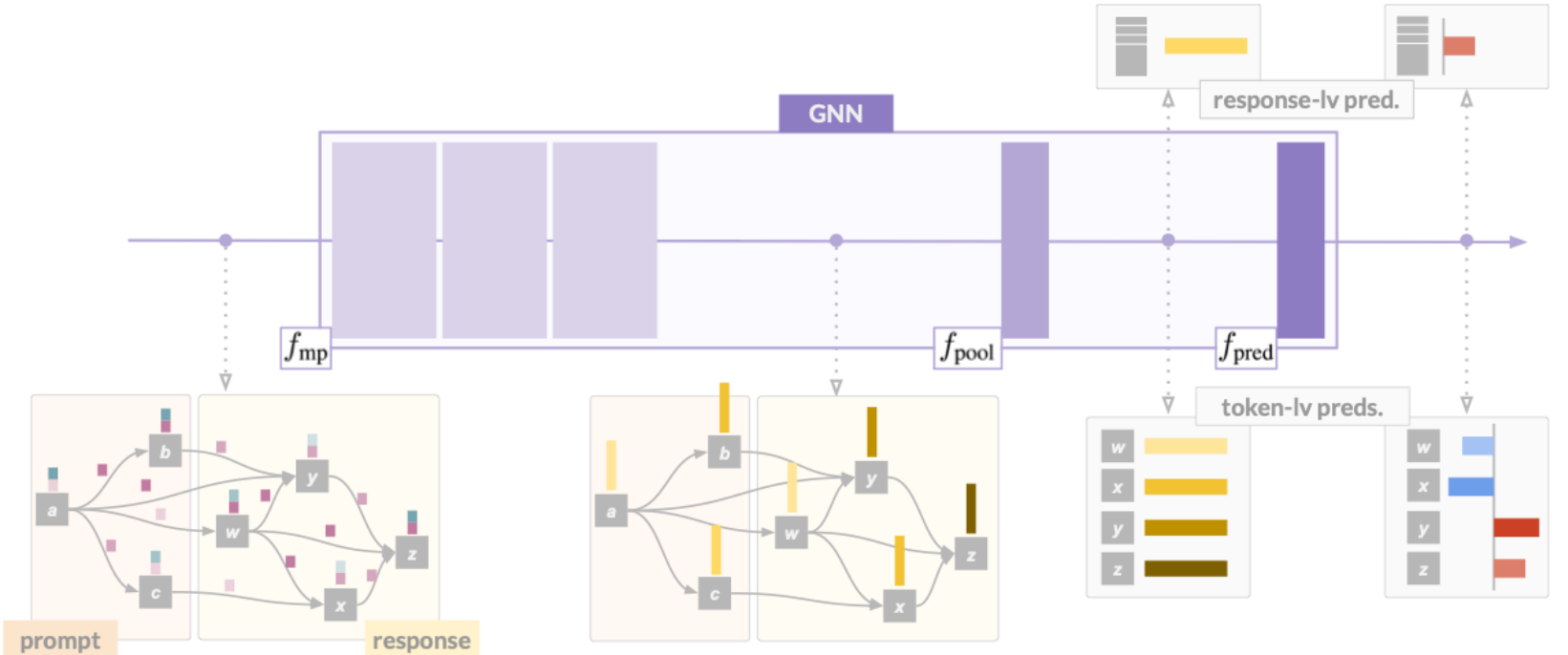


Example

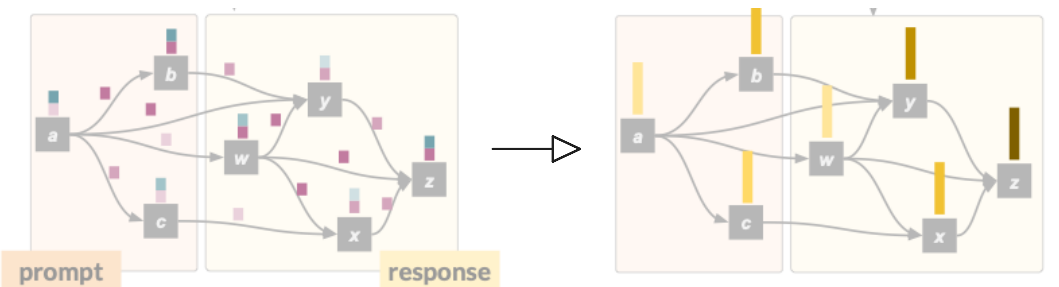
What kind of planet is Pluto? –> A dwarf planet



CHARM GNN

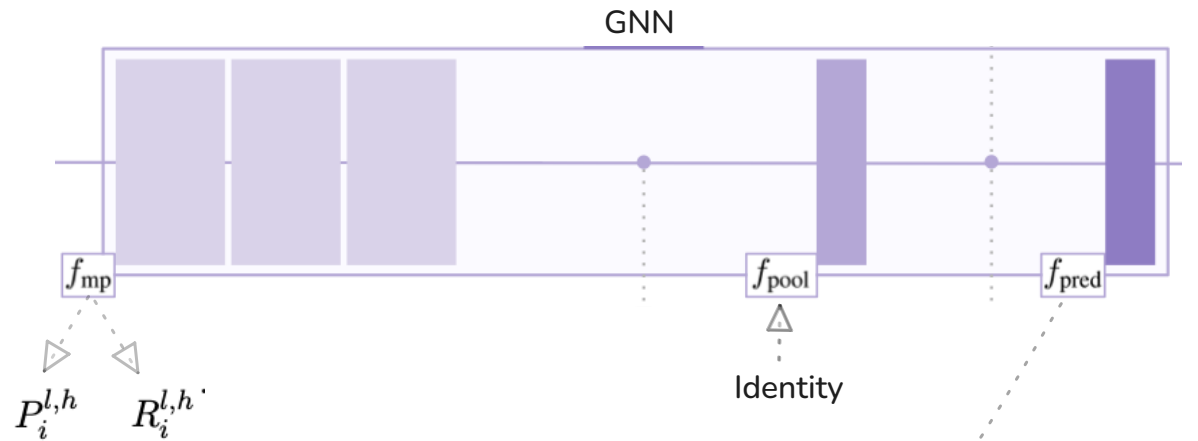

GNN
response-lv pred.
f_{mp}
f_{pool}
f_{pred}
token-lv preds.
a
b
c
w
x
y
z
prompt
response

CHARM Message Passing



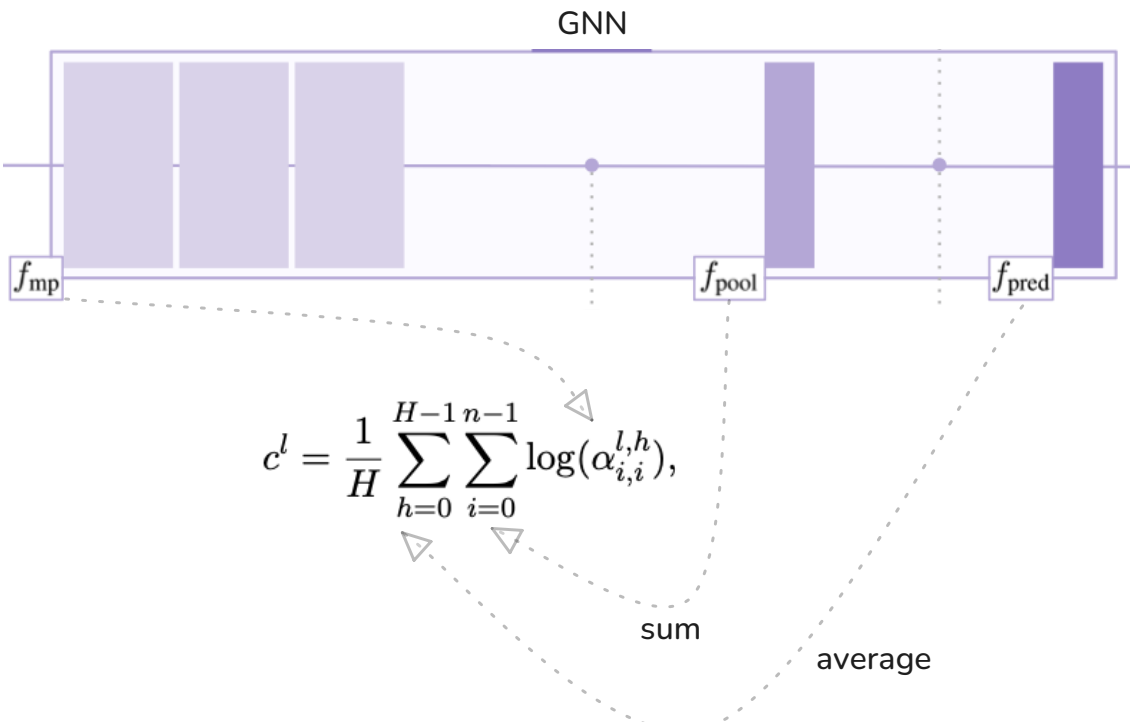
$$h_i^{(t+1)} = \mathrm{up}_t\left(h_i^{(t)}, \square_{j:\ (i,j)\in E} \mathrm{msg}_t\left(h_i^{(t)}, h_j^{(t)}, x^{\tau}_{E,(i,j)}, p_{i,j}\right)\right)$$

GNN Generalizes LookbackLens



$$\ell_i^{l,h} = \frac{P_i^{l,h}}{P_i^{l,h} + R_i^{l,h}}$$

GNN Generalizes LLM-Check



Contextual – Token Level

DATASET

NQ: Googled questions & Wikipedia answers

CNN: Text summarisation's of articles from CNN and Daily Mail

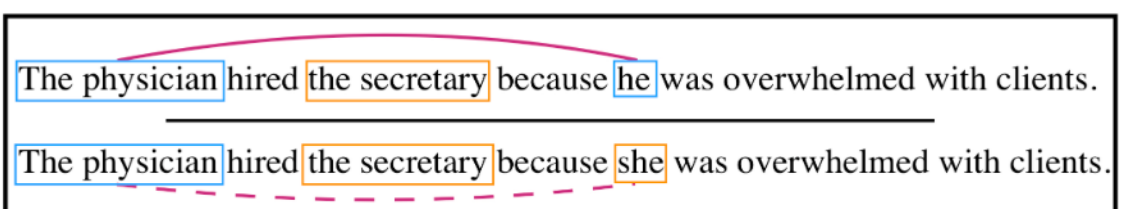
Non-Contextual Response Level

DATASET

Movies: Factual knowledge recalls

Math: Arithmetics

WinoBias: co-referencial bias



Zhao, Jieyu, et al, 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. arXiv, arXiv:1804.06876.

Contextual – Token Level

Table 1

	Method	NQ (Google Queries)		CNN (News Articles)	
		AUROC	AUPR	AUROC	AUPR
	Probas	49.8	16.2	54.4	8.2
	Act-24	<u>73.0</u>	<u>36.2</u>	71.3	<u>20.3</u>
	Act-28	71.6	34.6	70.1	18.4
	Act-32	67.4	28.6	67.7	15.4
	Lookback Lens	70.8	31.0	71.9	17.4
	Lookback Lens †	71.9	34.3	<u>74.4</u>	19.7
	Neigh-Avg(N)	66.0	24.5	70.1	14.9
	Neigh-Avg(E)	66.8	30.4	70.5	18.6
ours	CHARM (att)	**74.8**±0.6	**40.3**±1.7	**75.4**±0.2	**22.7**±0.4
ours	CHARM (att+act-24)	72.2±1.2	35.5±1.6	70.9±0.2	19.8±0.5

Non-Contextual Response Level

Table 2

	Method	Movies		Winobias		Math	
		AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
	Probas	58.6	81.6	64.5	20.0	54.5	57.4
	Act-24	77.0	90.4	76.6	37.8	77.7	77.5
	Act-28	77.0	90.4	73.9	34.3	78.1	77.8
	Act-32	76.3	90.2	72.7	35.3	76.6	77.9
	LLM-Chk-24	47.5	74.6	38.9	10.9	64.5	68.2
	LLM-Chk-28	51.1	76.7	41.6	11.4	65.5	69.2
	LLM-Chk-32	61.5	82.1	41.6	11.3	64.0	67.6
	LLM-Chk++-24	66.3	84.5	64.6	20.5	67.3	69.1
	LLM-Chk++-28	67.8	86.3	64.8	21.0	67.0	70.6
	LLM-Chk++-32	73.0	88.8	67.2	24.1	68.6	72.5
	LapEig	72.9	88.4	74.1	33.3	73.6	76.3
	Neigh-Avg(N)	78.6	91.2	63.8	23.0	77.4	79.2
	Neigh-Avg(E)	54.9	78.5	65.8	21.9	76.7	78.3
ours	CHARM (att)	**80.3**±0.2	**92.0**±0.1	70.4±0.7	29.1±1.0	76.5±1.1	79.7±0.5
ours	CHARM (att+act-24)	79.7±0.3	91.8±0.2	**77.8**±0.4	**39.8**±1.3	**80.8**±0.7	**83.1**±0.7

Ablation: Message Passing

Table 3

Method	CNN		Math	
	AUROC	AUPR	AUROC	AUPR
CHARM (no g.)	70.8 ±0.5	19.2 ±0.5	80.6 ±0.7	82.7 ±0.1
CHARM	**75.4** ±0.2	**22.7** ±0.4	**81.7** ±0.2	**83.8** ±0.3

Ablation: Attention Threshold

Table 4

τ	Num. Edges	Sparsity	Mem. (MB)	AUPR
0.5	1,118.80	0.993	$22.99_{\pm 3.71}$	$38.4_{\pm 0.4}$
0.1	7,458.67	0.952	$60.44_{\pm 11.79}$	$41.0_{\pm 1.2}$
0.05	14,884.44	0.906	$104.15_{\pm 23.05}$	$40.3_{\pm 1.7}$
0.01	58,998.88	0.645	$363.02_{\pm 98.39}$	$40.3_{\pm 0.9}$
0.001	19,7784.82	0.026	$1177.20_{\pm 523.61}$	$40.1_{\pm 0.0}$

Ablation: Generalization

Table 5.1

Method	CNN → NQ	
	AUROC	AUPR
LkbLens †	62.0	26.5
Act-24	63.8	29.7
CHARM	**65.5** ±0.14	**31.6** ±0.10

News Article Summarization –> Google Queries

Ablation: Generalization

Table 5.2

Method	NQ → CNN	
	AUROC	AUPR
LkbLens †	**68.6**	**14.9**
Act-24	63.4	11.3
CHARM	64.1 ±1.1	12.0 ±0.98

Google Queries –> News Article Summarization

My Opinion

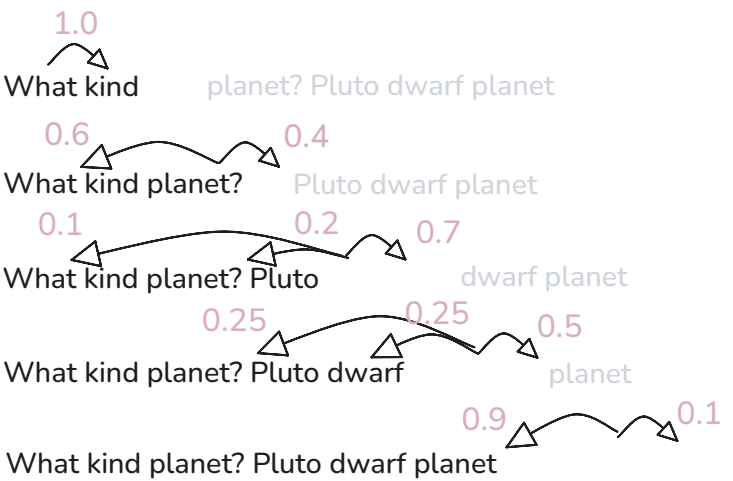
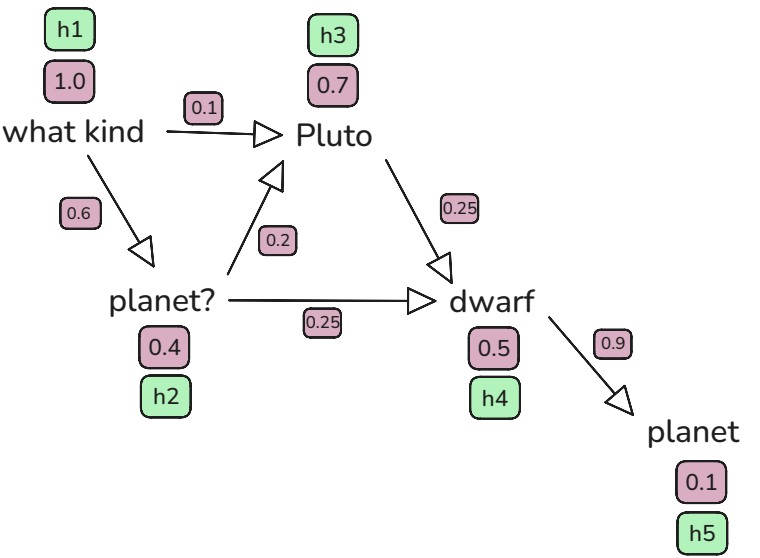
STRENGTHS

* Framework
* unifies token/response levels
* enables/ unleashes deep learning
* Generalizes prior work
* Attention helps detect contextual HD

WEAKNESSES

* 'Simple' categorisation of Hallucinations
* Multiple layers not reported
* ..... [1]Code will be released soon.

THANKS!
QUESTIONS?

Extra Slides 1

What kind of planet is Pluto? –> A dwarf planet