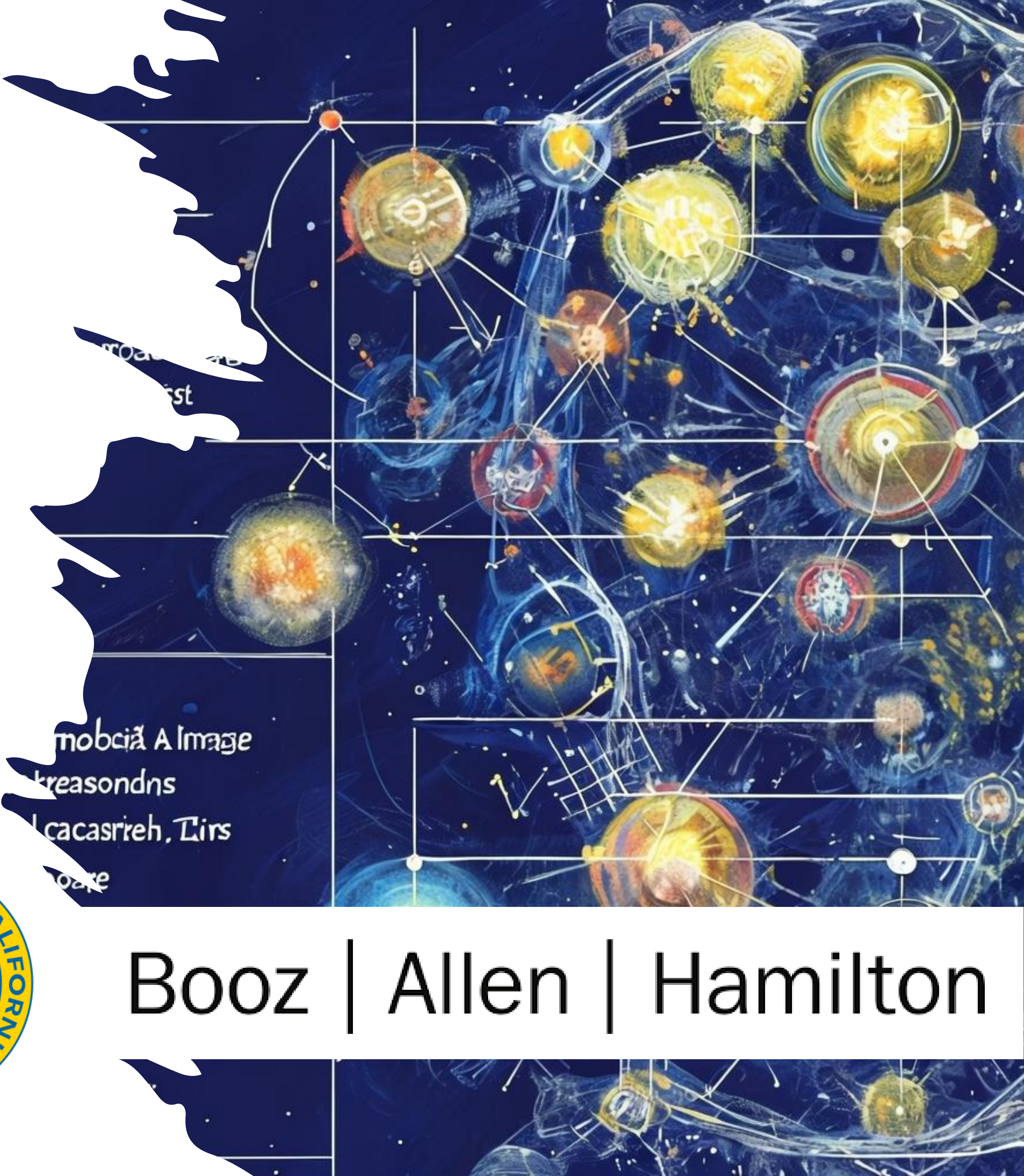
aaai Tutorial 2025

NeuroSymbolic AI for Grounding, Instructibility, and Explainability



UMBC

amazon

UNIVERSITY OF South Carolina

THE UNIVERSITY OF CALIFORNIA IRVINE

Booz | Allen | Hamilton

Spread the Word

Making LLMs Explainable, Grounded, and Instructible

Explainability

#LLMExplainability #InterpretableAI #ExplainableLLM #TransparentAI

Grounding

#LLMGrounding #GroundedLLM #FunctionalGrounding #SymbolicGrounding #AISymbolicReasoning #NeuroSymbolicLLM #FunctionalLLM

Instructibility

#LLMInstructibility #LLMSafety #AIInstructability #LLMTutorial

As of June 2024

Time

Source: Gartner

Gartner

Forthcoming Summer 2025 from Cambridge!

Knowledge-Infused Learning: Neurosymbolic AI for Explainability, Interpretability, and Safety

Manas Gaur, Amit P. Sheth

Focus of Tutorial

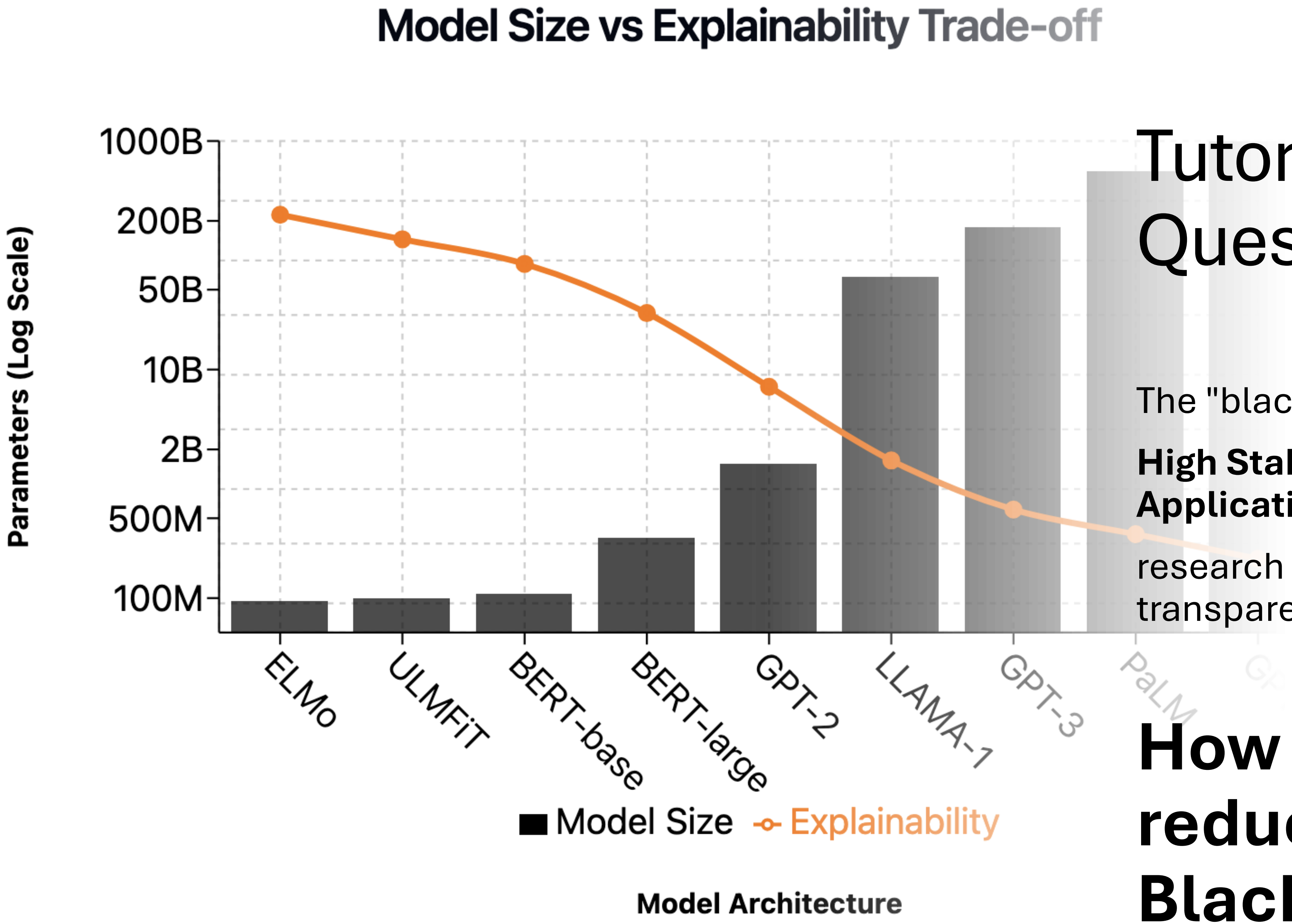
NeuroSymbolic AI and Instructible AI

Vector Symbolic Architectures

Grounding with Retrieval Augmented Generation

Explainability with Knowledge-infused Learning

OpenCHA: Domain Knowledge-driven LLM-based Conversational Agent for Health



Tutorial's Central Question

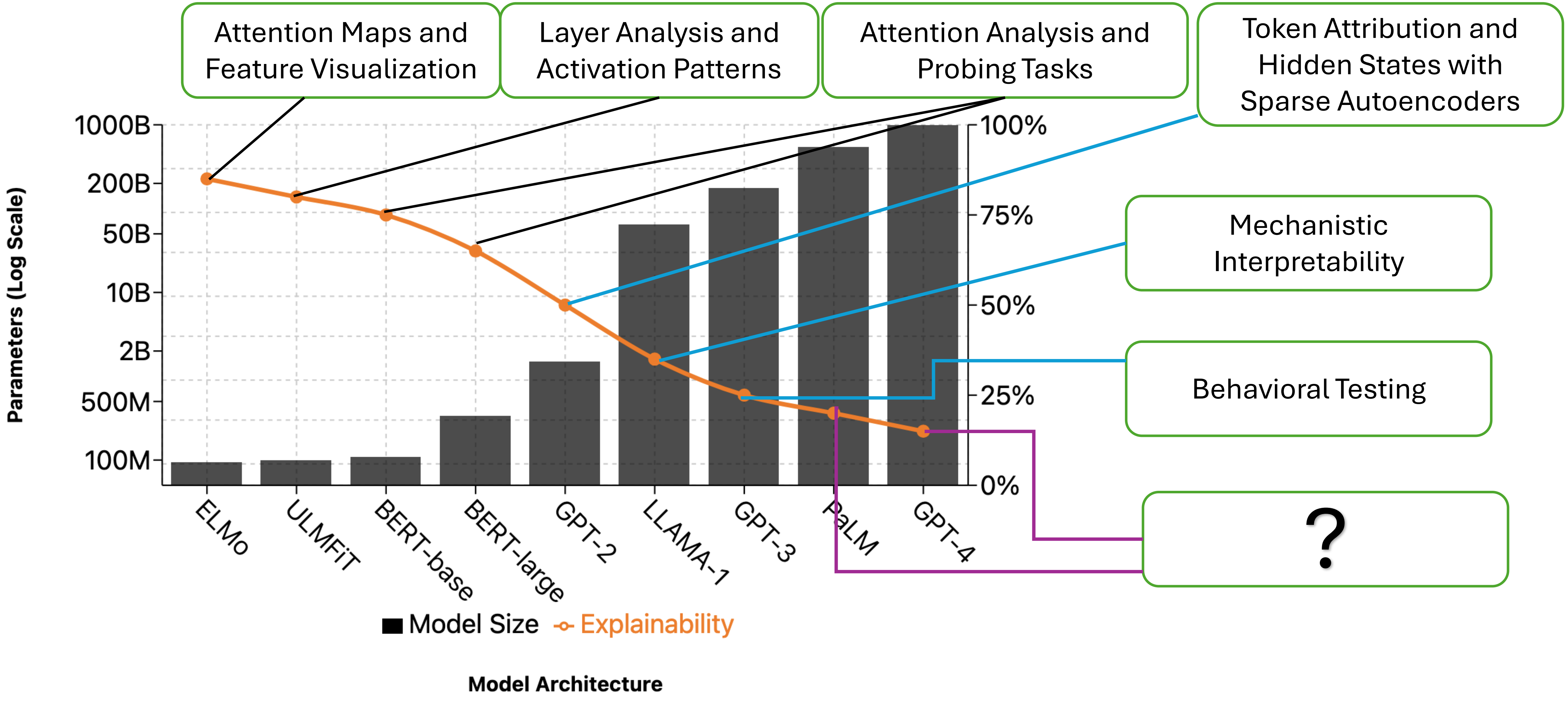
The "black box" nature of AI systems in

High Stakes Decision-Making Application

research has raised concerns about transparency and reproducibility.

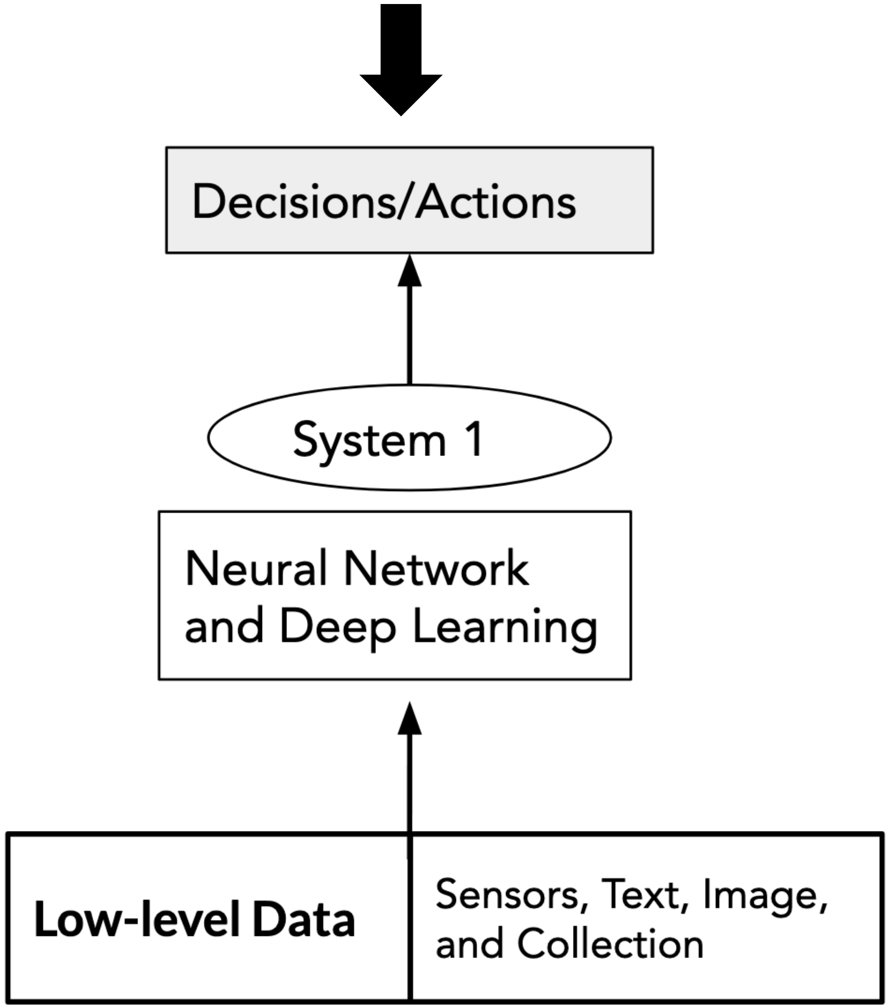
How can we go about reducing Blackboxness??

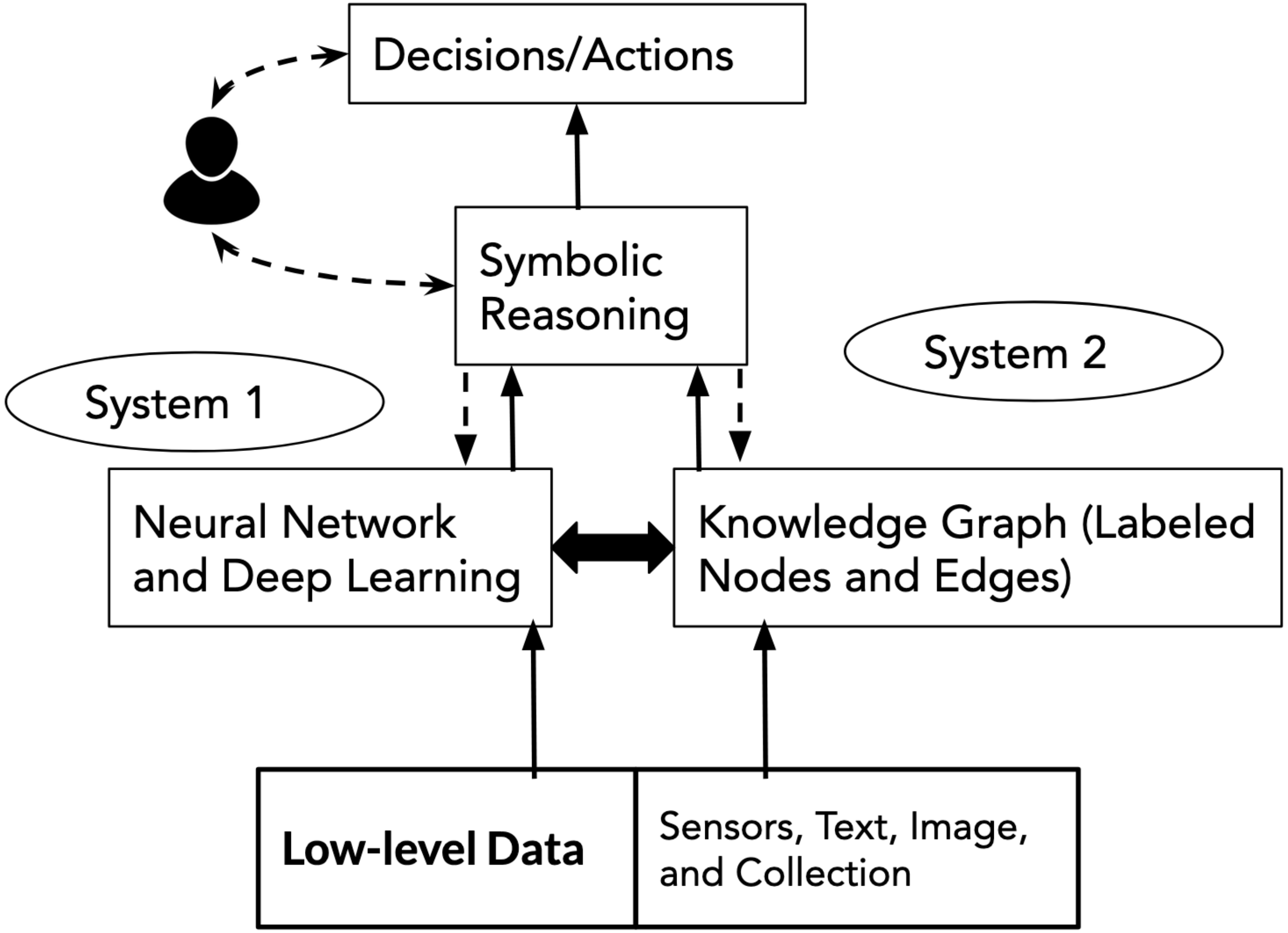
Note: Explainability scores are approximate and based on available interpretation techniqu



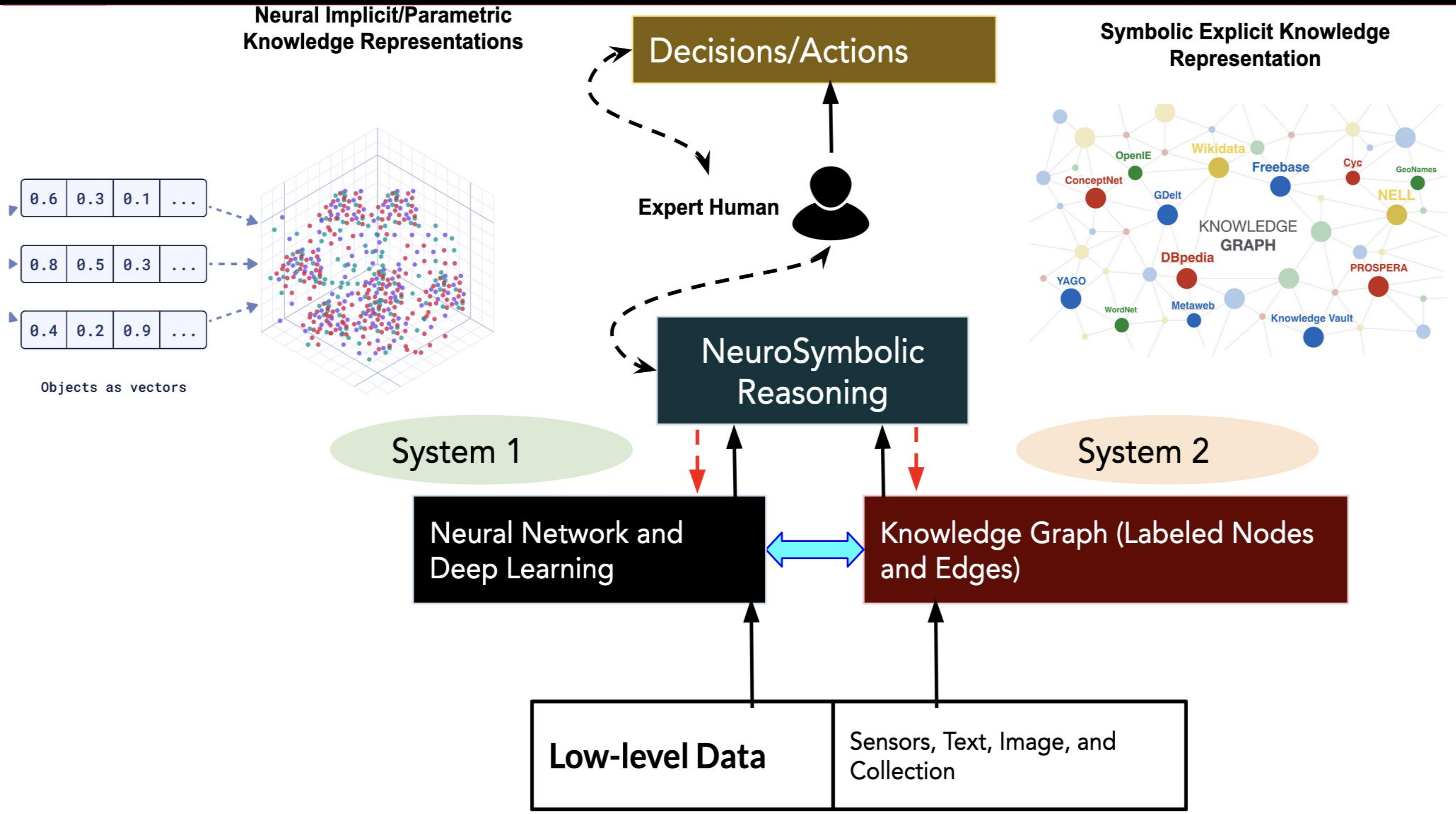
Note: Explainability scores are approximate and based on available interpretation techniques for each

Statistical AI is a Blackbox ➡ NeuroSymbolic AI





Why NeuroSymbolic AI



Amit Sheth, Kaushik Roy, Manas Gaur, Neurosymbolic Artificial Intelligence (Why, What, and How), IEEE Intelligent Systems, 38 (3), May-June 2023

Instructability

➔ The capability of AI systems to be taught and guided by humans to cause intentional behaviours.

➔ **Features:**
- Skill Acquisition
- Knowledge-Gap management
- Human-AI Interaction
  - Explainability
  - Interpretability
  - Observability

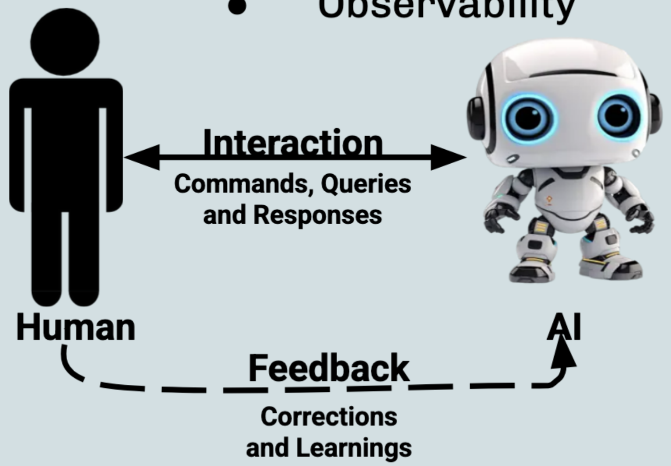


Image by oval on clker.com
Image by pngtree.com

Grounding

➔ The process of establishing meaningful connections between AI representations and the real world, ensuring AI systems understand and interact with their environment effectively.

→ Features:
- Symbolic Grounding
- Functional Grounding

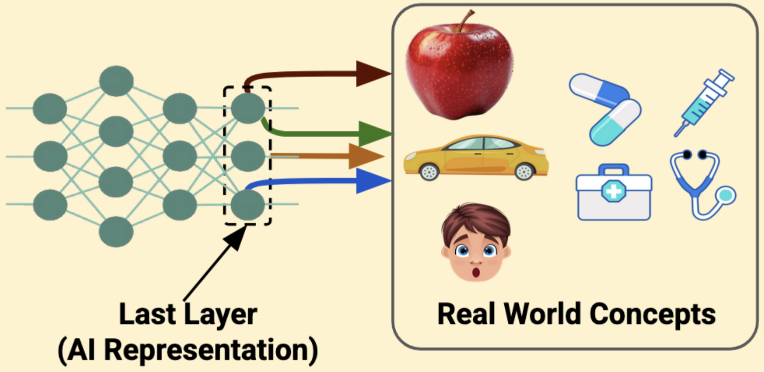
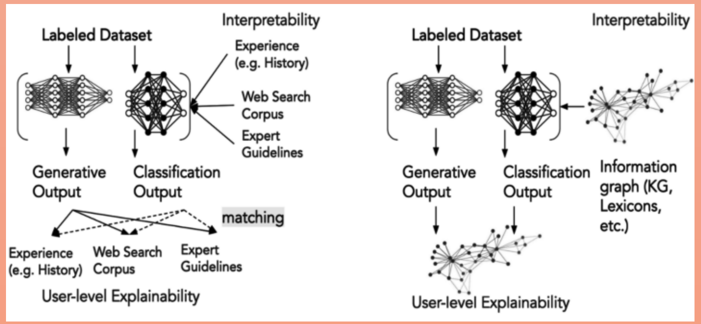


Image by pngtree.com
Image by medium
Image by emiltimplaru, santima.studio on vecteezy

Explainability

→ Explanations that use terms and connections specific to a particular field or industry are more useful than general words that don't help people take action.

→ Features:
- Local and Global Explanations
- Causality
- User-Appropriate Explanations



NeuroSymbolic AI

Neuro-symbolic AI techniques **incorporate broader forms of knowledge** (**lexical, domain-specific, common-sense, and constraint-based)** into addressing limitations of either symbolic or statistical AI approaches, such as **model interpretations and user-level explanations.** **Compared to powerful statistical AI that exploits data, NeSy benefits from data and knowledge.**

Neurosymbolic AI in the Era of Large Language Models

Copilots for Health, Manufacturing

Implicit Personalization in Domain-specific Conversations

Workflows-guided Safe Decision Making

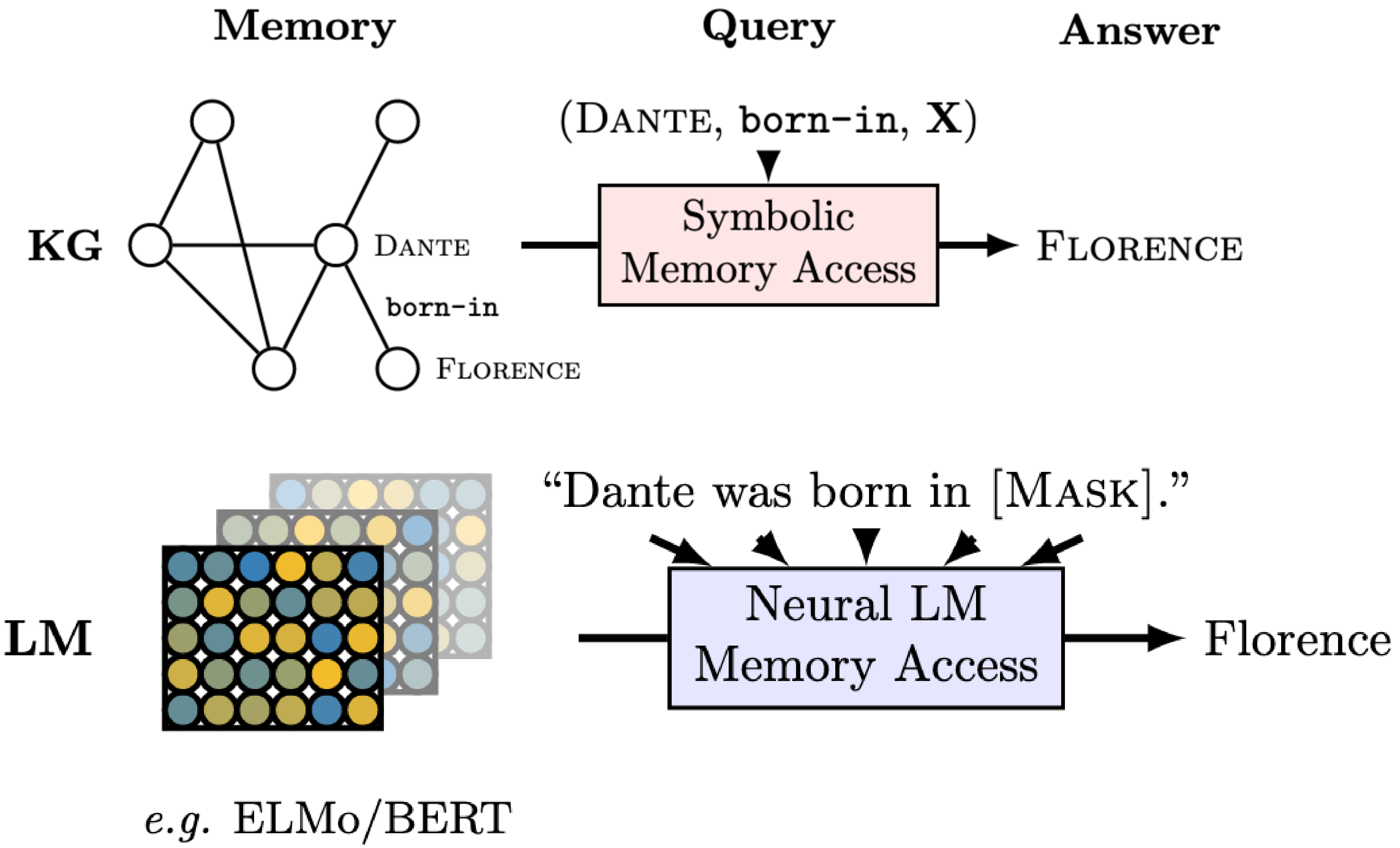
Explainable and Attributable Reasoning on Narratives

Grounded Debiasing using Rules and Domain Knowledge (No additional Data)

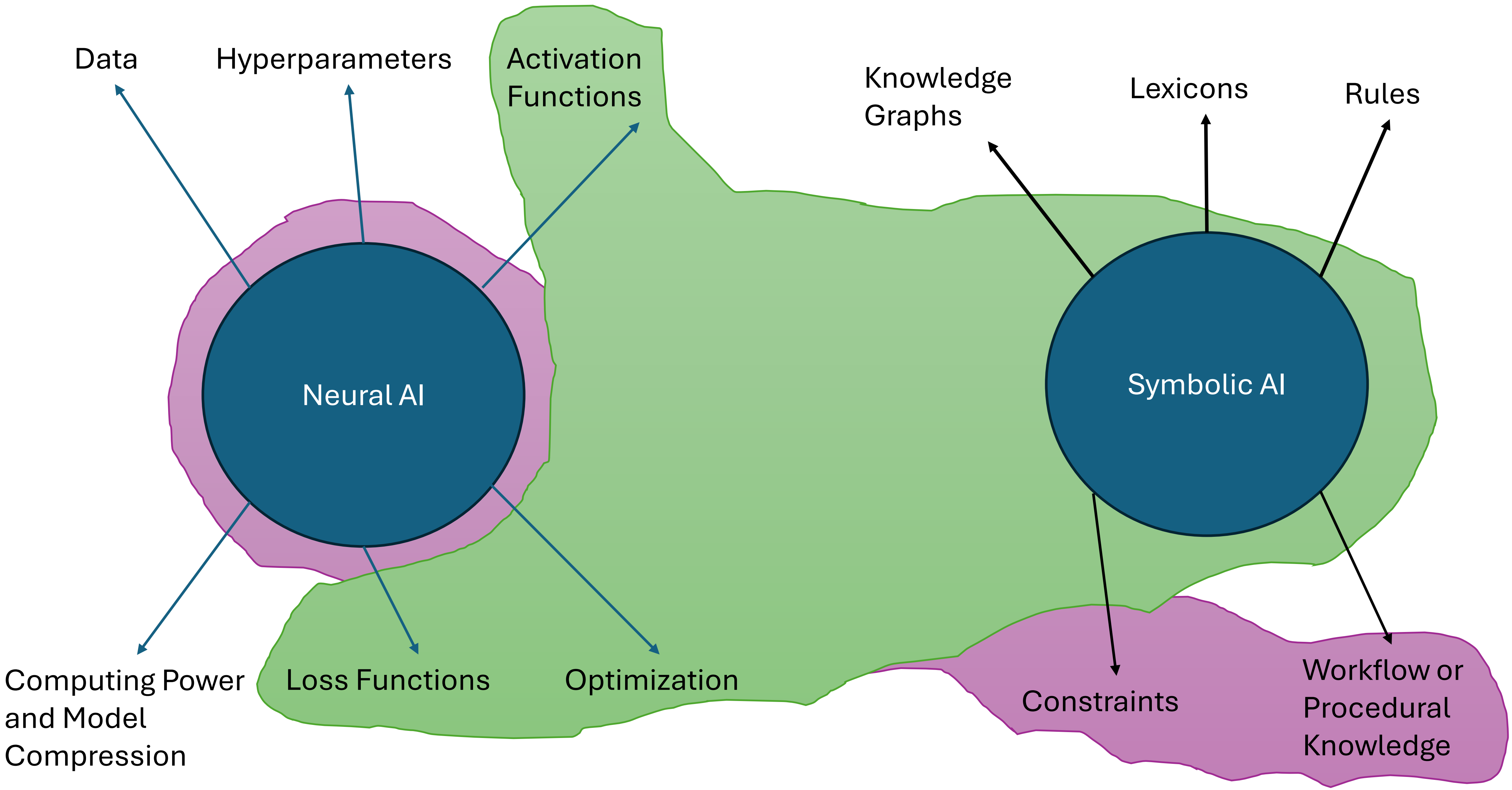
Social Welfare Optimized Ensemble of LLMs

Applications of NeuroSymbolic AI with a focus on Grounding, Explainability, and Instructibility (EGI)

NeuroSymbolic in Machine Learning and Natural Language Processing



Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., & Miller, A. (2019, November). Language Models as Knowledge Bases?. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*


Data
Hyperparameters
Activation Functions
Knowledge Graphs
Lexicons
Rules
Neural AI
Symbolic AI
Computing Power and Model Compression
Loss Functions
Optimization
Constraints
Workflow or Procedural Knowledge

Benchmarking Example

WellDunn: On the Robustness and Explainability of Language Models and Large Language Models in Identifying Wellness Dimensions,
In BlackboxNLP @ EMNLP 2024

Wellness Dimension

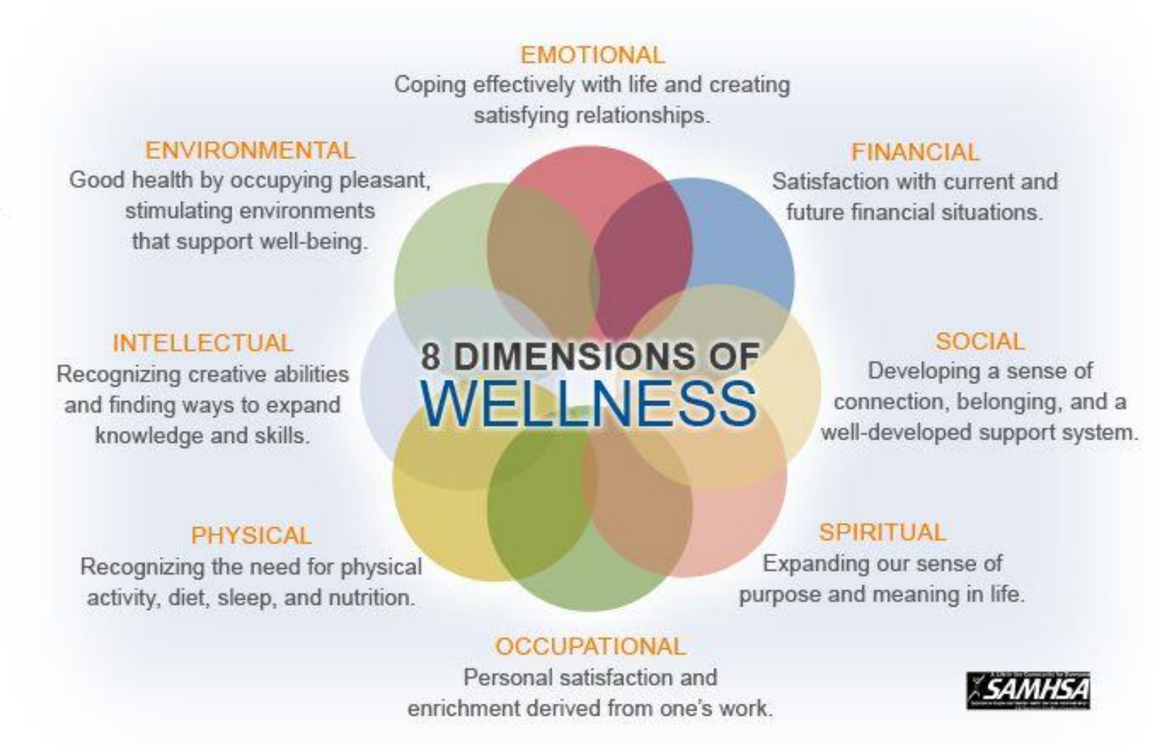
Source: Adapted from Swarbrick, M. (2006). A Wellness Approach. *Psychiatric Rehabilitation Journal, 29*(4), 311–314.

Substance Abuse and Mental Health Services Administration
SAMHSA
www.samhsa.gov

DEPARTMENT OF HEALTH & HUMAN SERVICES · USA

Wellness Dimension

Wellness Dimension Definitions and Questionnaire



Substance Abuse and Mental Health Services Administration
SAMHSA
www.samhsa.gov

DEPARTMENT OF HEALTH & HUMAN SERVICES · USA

https://store.samhsa.gov/sites/default/files/sma16-4958.pdf

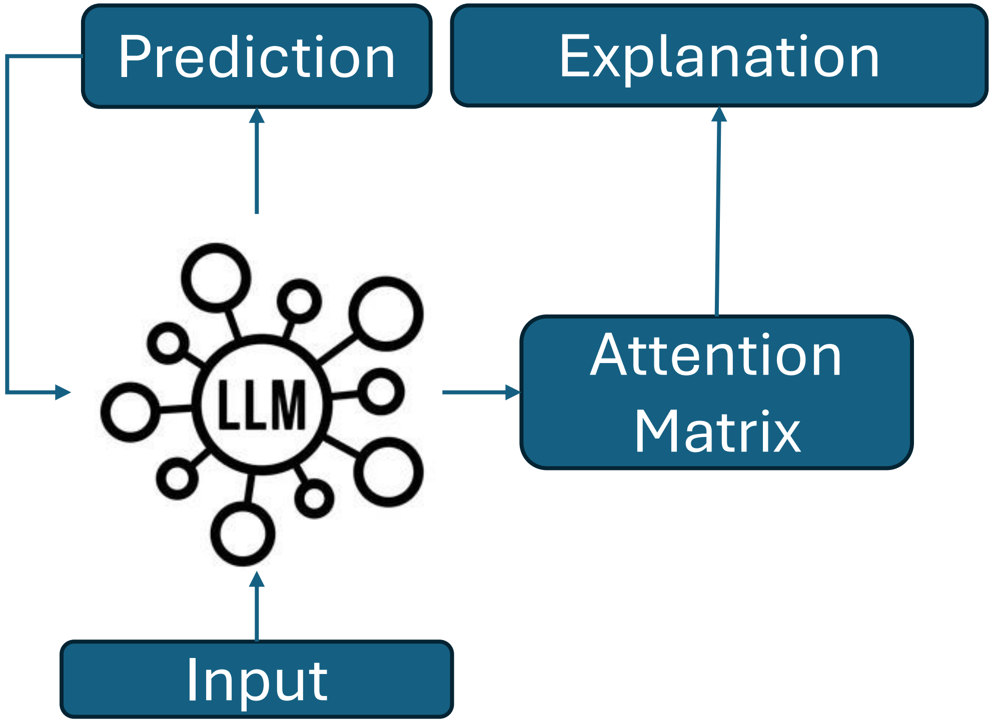
MultiWD and WellXplain Datasets

Content worth 4000 users

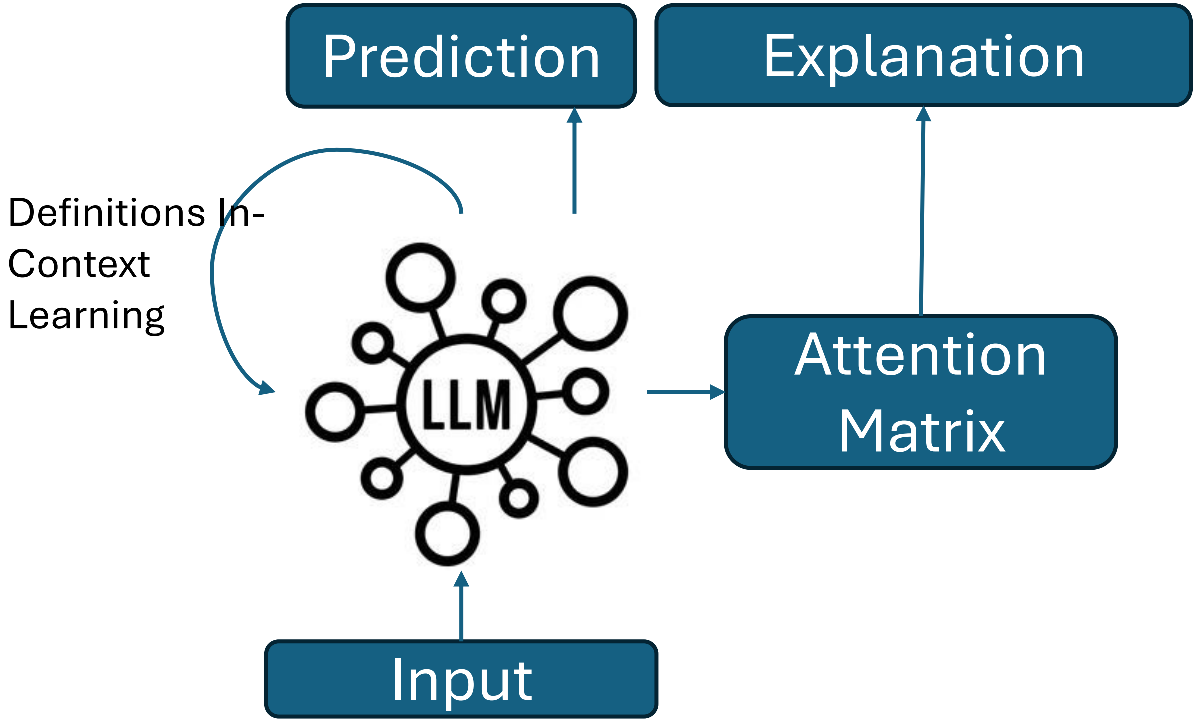
6 Wellness Dimensions

- Physical
- Intellectual
- Vocational
- Social
- Spiritual
- Emotional

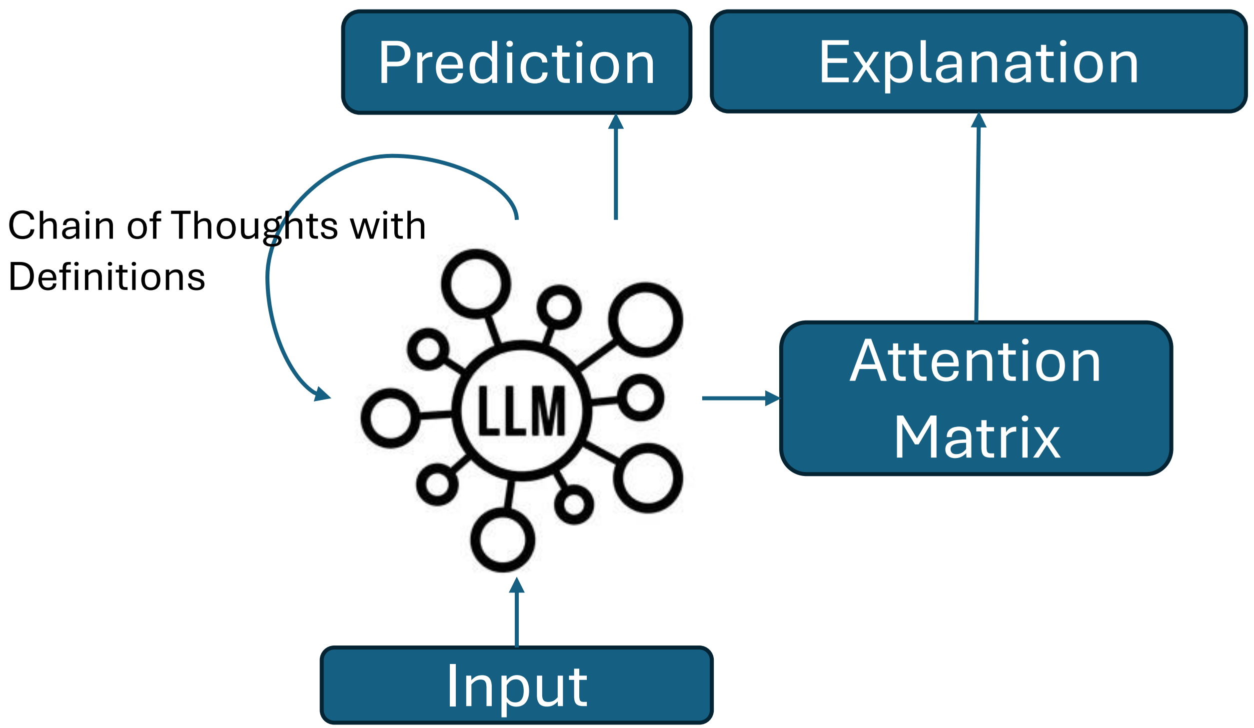
Clinical expert explanations



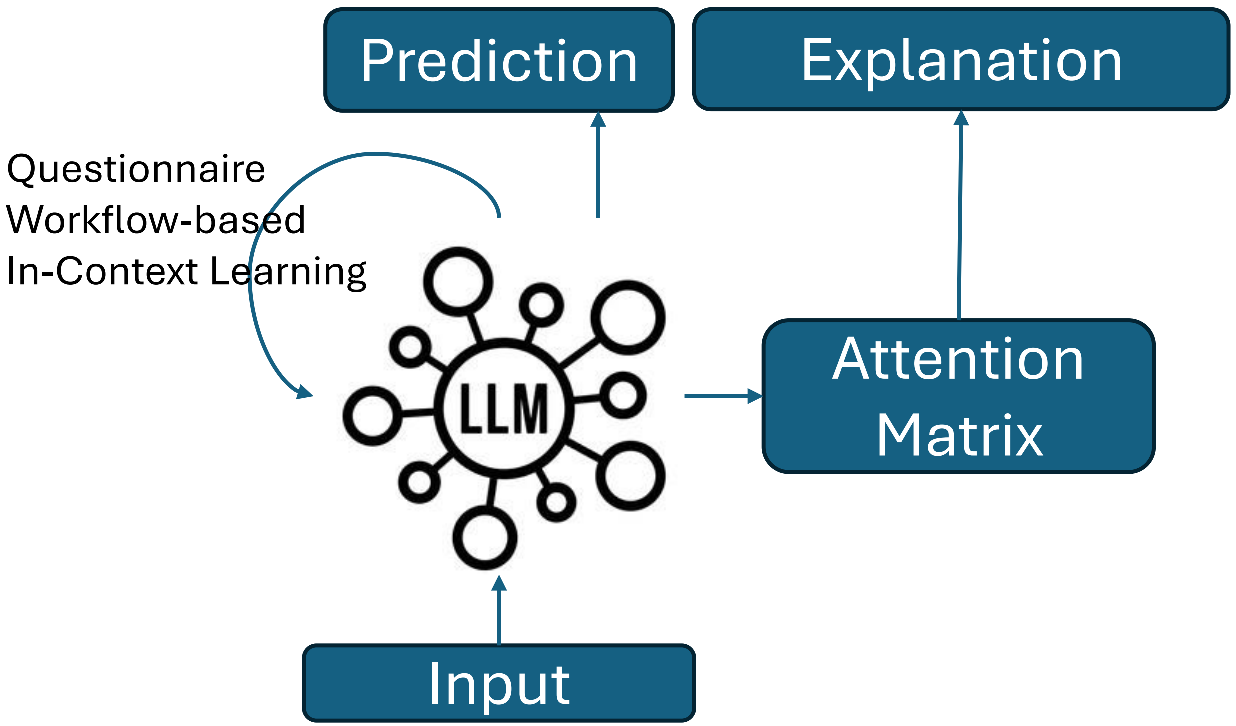
Design 1



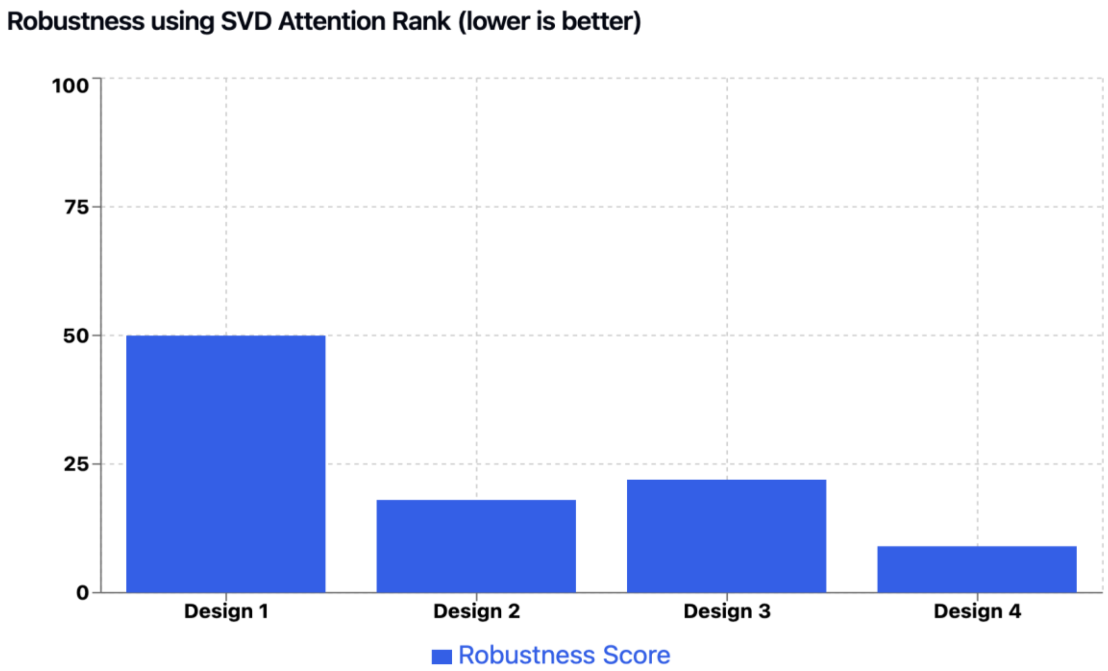
Design 2

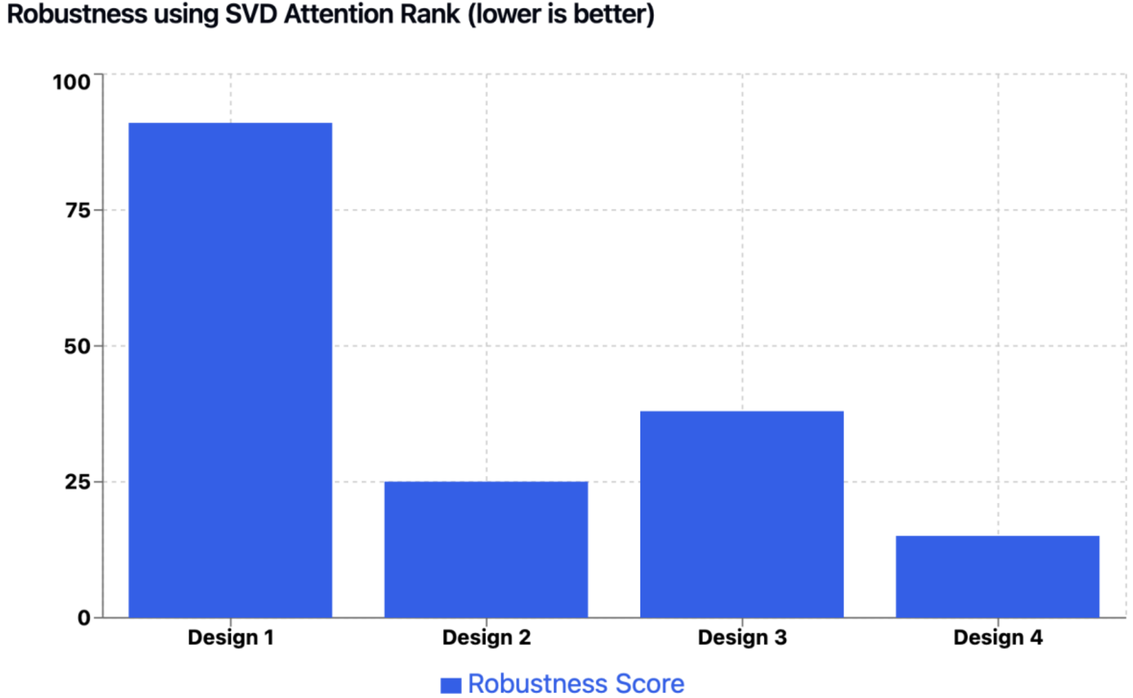


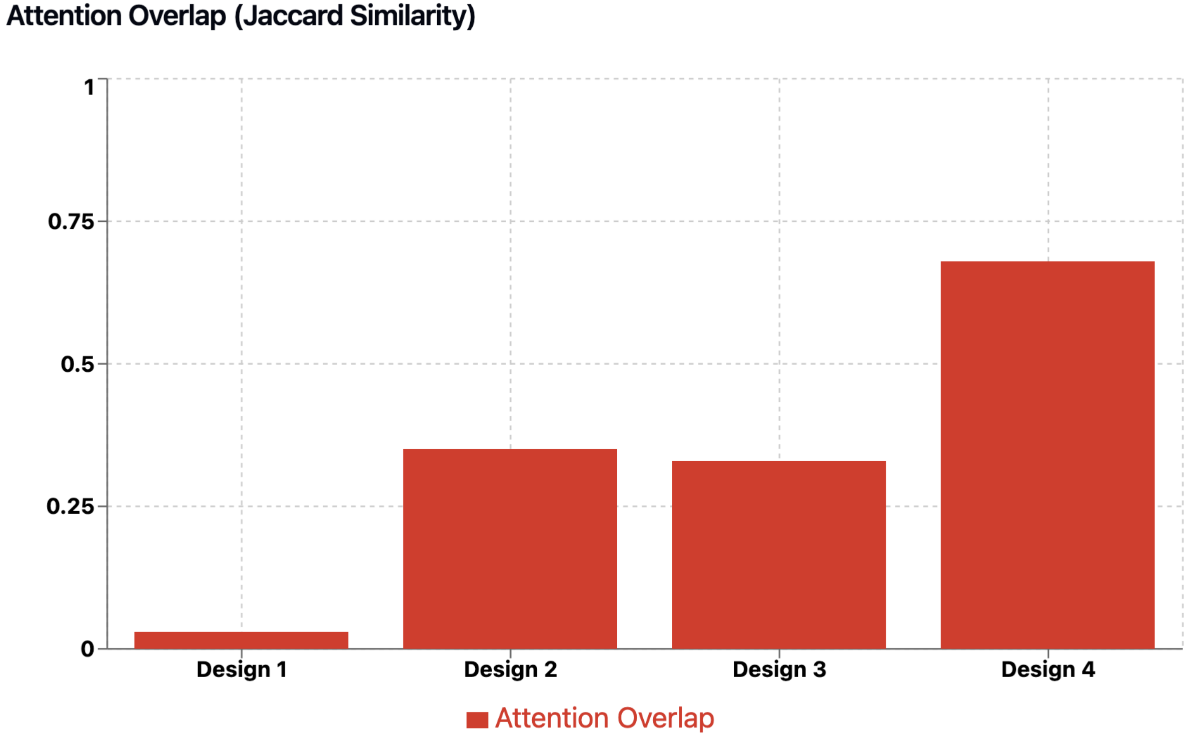
Design 3

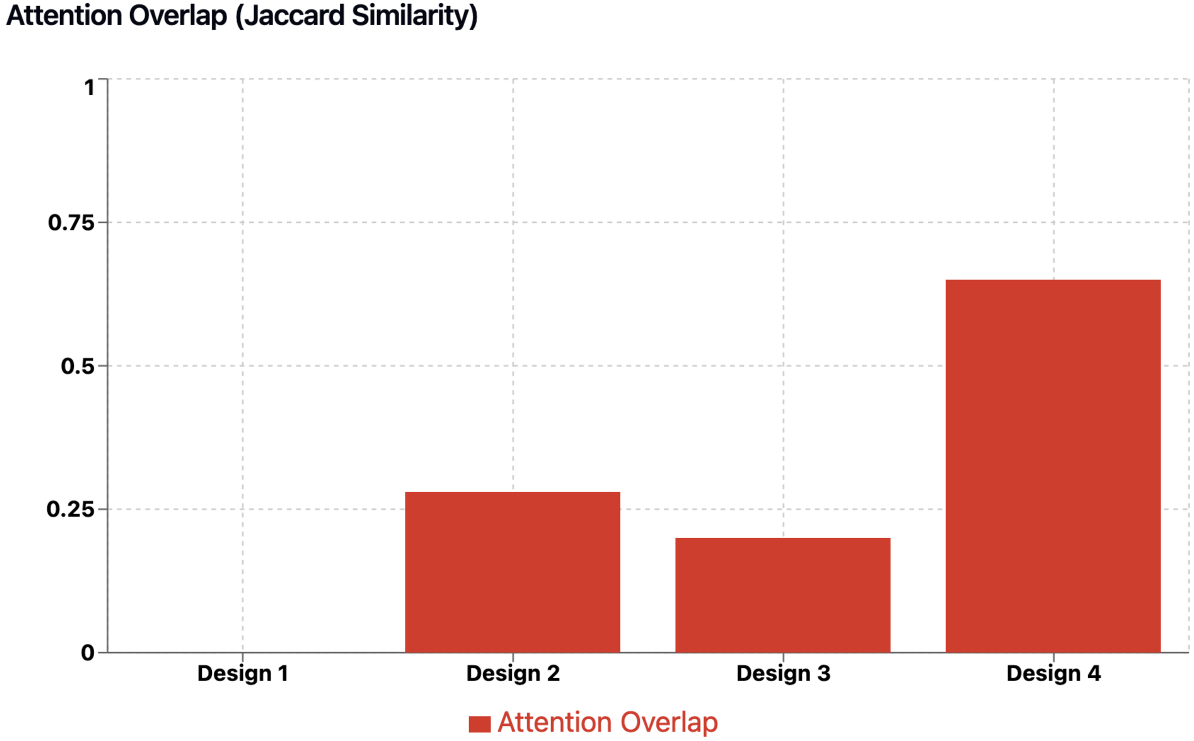


Design 4

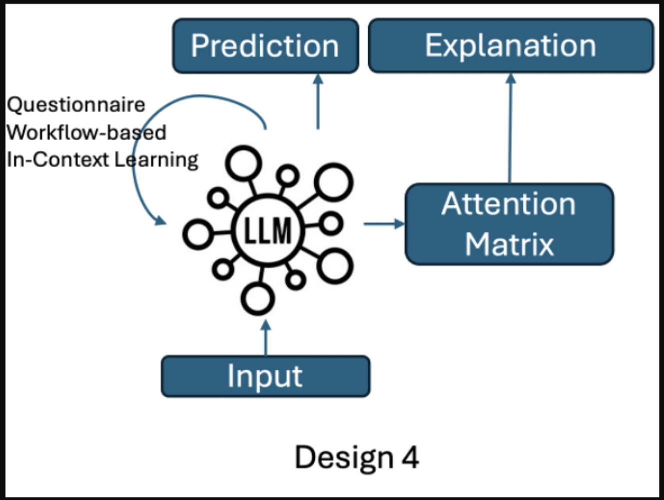

Robustness using SVD Attention Rank (lower is better)
100
75
50
25
0
Design 1
Design 2
Design 3
Design 4
Robustness Score

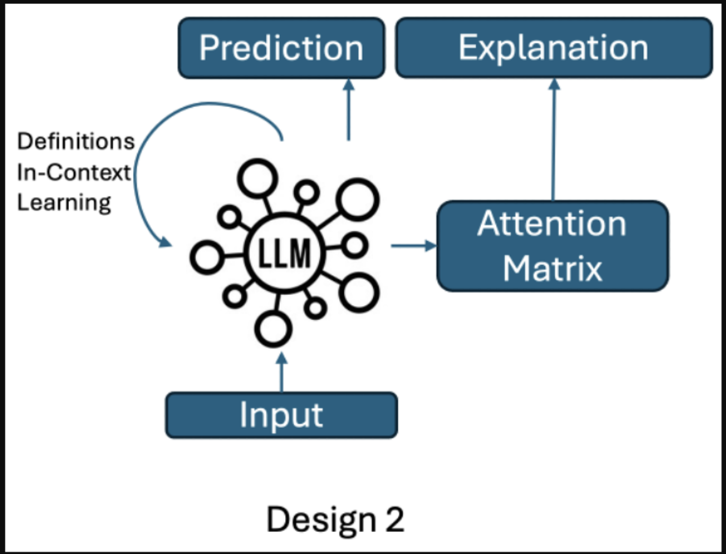

Robustness using SVD Attention Rank (lower is better)
100
75
50
25
0
Design 1
Design 2
Design 3
Design 4
Robustness Score


Attention Overlap (Jaccard Similarity)
1
0.75
0.5
0.25
0
Design 1
Design 2
Design 3
Design 4
Attention Overlap


Attention Overlap (Jaccard Similarity)
1
0.75
0.5
0.25
0
Design 1
Design 2
Design 3
Design 4
Attention Overlap

Hybridized Architectures: NeuroSymbolic AI





Process Knowledge-infused Learning

Domain Data

Raw data from specific domain and processes

Process Knowledge

Expert-defined rules, constraints, and process sequ

Knowledge Embedding

Transform process knowledge into neural represent

Infused Learning

Joint training with data and embedded knowledge

Key Benefits:

- Enhanced model interpretability through explicit process knowledge
- Improved generalization by incorporating domain expertise
- Better constraint satisfaction in predictions
- Reduced data requirements through knowledge guidance

Simple Text Classification

I am really struggling with my bisexuality, which is causing chaos in my relationship with a girl. Being a fan of the LGBTQ community, I am equal to worthless to her. I'm now starting to get drunk because I can't cope with the obsessive, intrusive thoughts I need to get out of my head.
Don't want to live anymore. Sexually assault, ignorant family members, and my never-ending loneliness brights up my path to death.
I do have the potential to live a decent life, but not with people who abandon me. Hopelessness and feelings of betrayal have turned my nights into days. I am developing insomnia because of my restlessness.
I just can't take it anymore. Been abandoned yet again by someone I cared about. I've been diagnosed with borderline for a while, and I'm just going to isolate myself and sleep forever.

Y = Suicide Ideation

Process Knowledge Infusion is better form than data-driven Classification

Simple Text Classification

I am really struggling with my bisexuality, which is causing chaos in my relationship with a girl. Being a fan of the LGBTQ community, I am equal to worthless to her. I'm now starting to get drunk because I can't cope with the obsessive, intrusive thoughts I need to get out of my head.
Don't want to live anymore. Sexually assault, ignorant family members, and my never-ending loneliness brights up my path to death.
I do have the potential to live a decent life, but not with people who abandon me. Hopelessness and feelings of betrayal have turned my nights into days. I am developing insomnia because of my restlessness.
I just can't take it anymore. Been abandoned yet again by someone I cared about. I've been diagnosed with borderline for a while, and I'm just going to isolate myself and sleep forever.

Y = Suicide Ideation

Process Knowledge-based Classification

Has the subject wished he was dead or wished he could go to sleep and not wake up?

YES

Has the subject had any thoughts of killing himself?

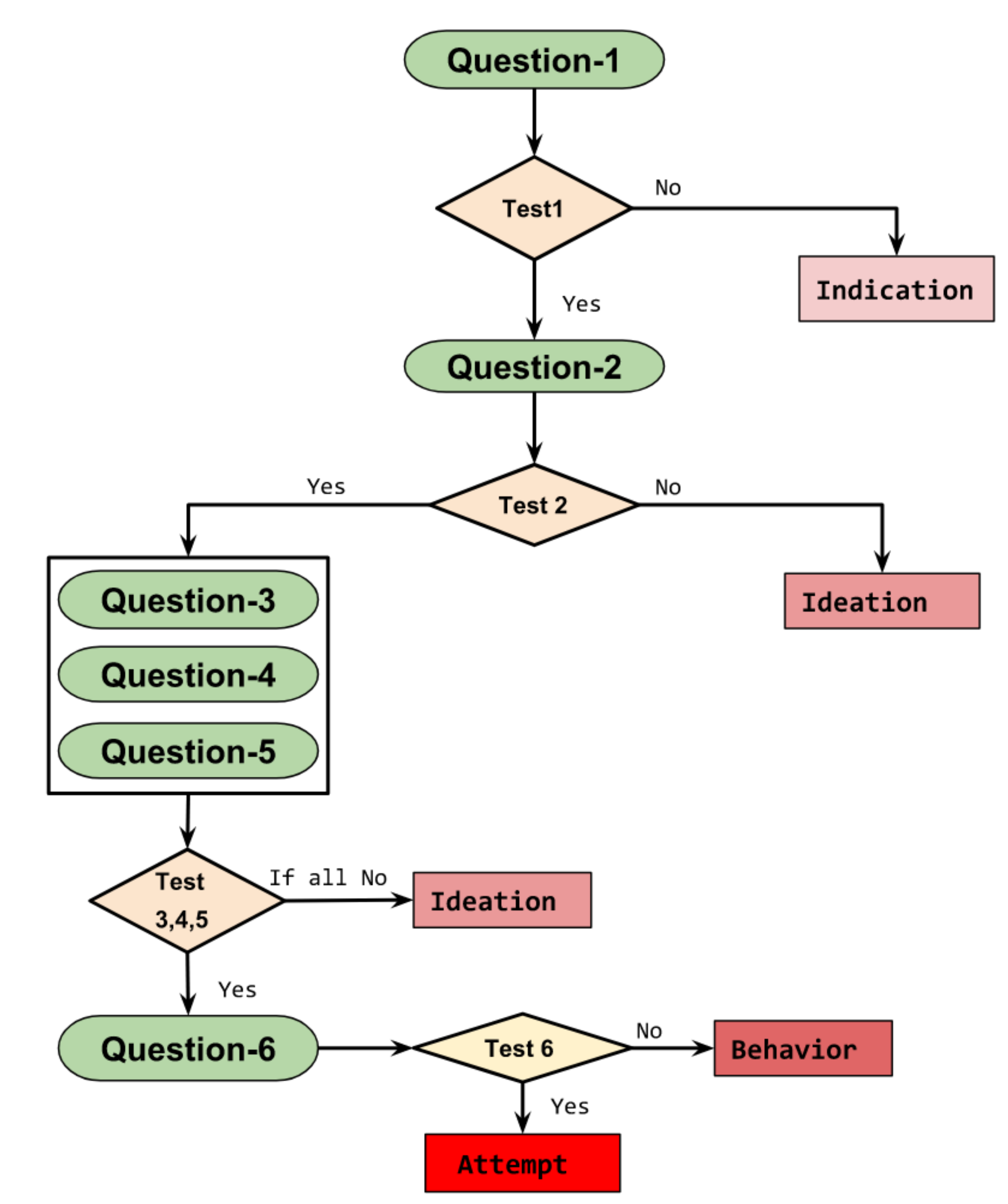
YES

Has the subject been thinking about how he might do this?

NO

Has the subject has these thoughts and some intentions of acting on them?

NO

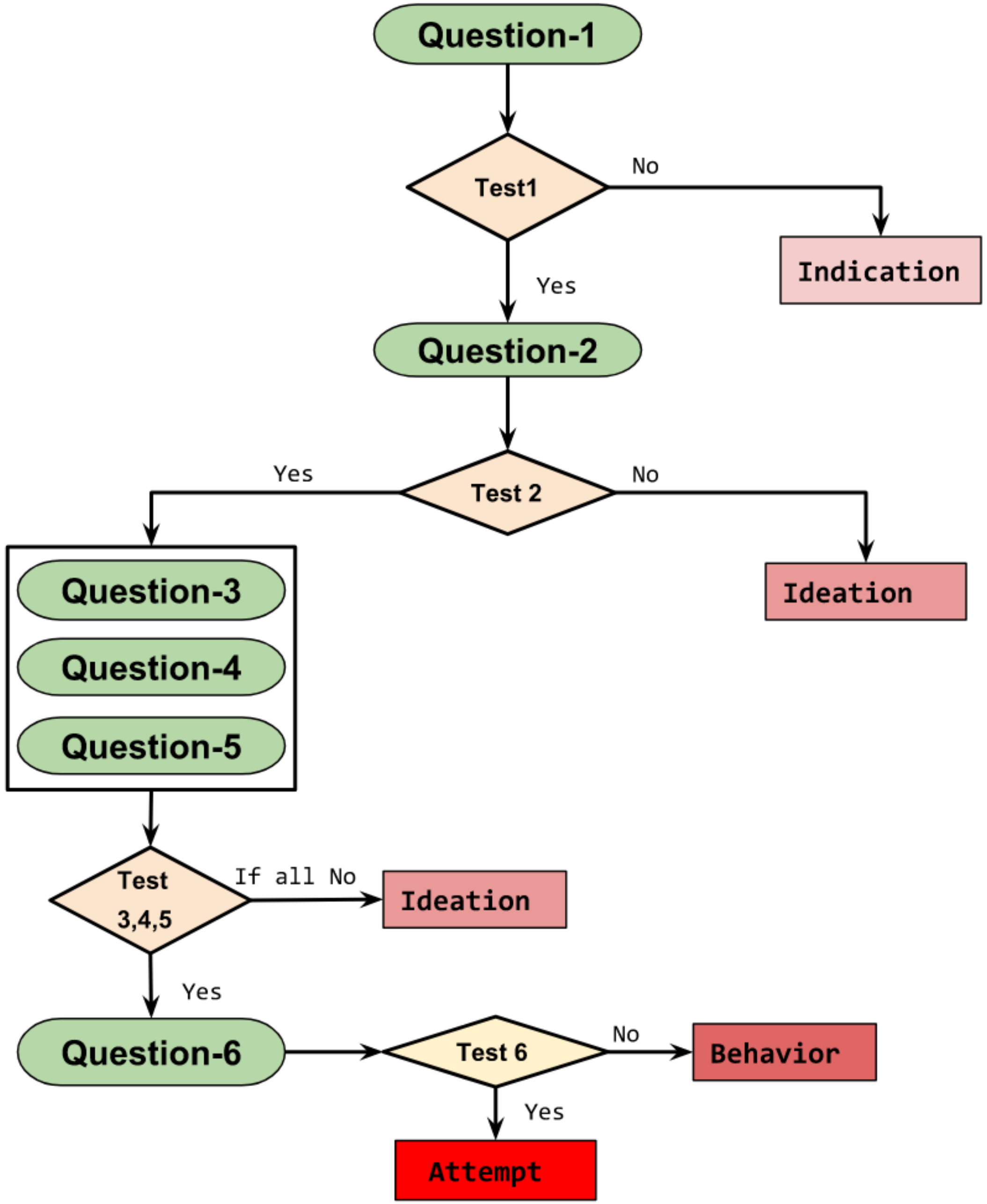


Simple Text Classification

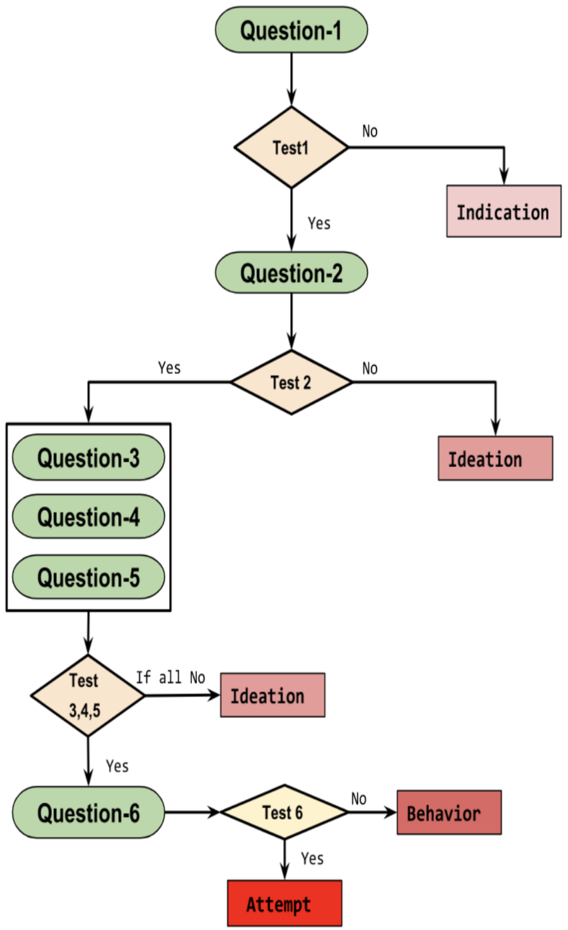
I am really struggling with my bisexuality, which is causing chaos in my relationship with a girl. Being a fan of the LGBTQ community, I am equal to worthless to her. I'm now starting to get drunk because I can't cope with the obsessive, intrusive thoughts I need to get out of my head.
Don't want to live anymore. Sexually assault, ignorant family members, and my never-ending loneliness brights up my path to death.
I do have the potential to live a decent life, but not with people who abandon me. Hopelessness and feelings of betrayal have turned my nights into days. I am developing insomnia because of my restlessness.
I just can't take it anymore. Been abandoned yet again by someone I cared about. I've been diagnosed with borderline for a while, and I'm just going to isolate myself and sleep forever.

Y = Suicide Ideation

SUICIDE IDEATION DEFINITIONS AND PROMPTS	**Past month**	
Ask questions that are bolded and <u>underlined</u>.	**YES**	**NO**
Ask Questions 1 and 2		
1) Wish to be Dead: Person endorses thoughts about a wish to be dead or not alive anymore, or wish to fall asleep and not wake up. ***<u>Have you wished you were dead or wished you could go to sleep and not wake up?</u>***		
2) Suicidal Thoughts: General non-specific thoughts of wanting to end one's life/commit suicide, "*I've thought about killing myself*" without general thoughts of ways to kill oneself/associated methods, intent, or plan. ***<u>Have you actually had any thoughts of killing yourself?</u>***		
If YES to 2, ask questions 3, 4, 5, and 6. If NO to 2, go directly to question 6.		
3) Suicidal Thoughts with Method (without Specific Plan or Intent to Act): Person endorses thoughts of suicide and has thought of a least one method during the assessment period. This is different than a specific plan with time, place or method details worked out. "*I thought about taking an overdose but I never made a specific plan as to when where or how I would actually do it….and I would never go through with it.*" ***<u>Have you been thinking about how you might kill yourself?</u>***		
4) Suicidal Intent (without Specific Plan): Active suicidal thoughts of killing oneself and patient reports having <u>some intent to act on such thoughts</u>, as opposed to "*I have the thoughts but I definitely will not do anything about them.*" ***<u>Have you had these thoughts and had some intention of acting on them?</u>***		
5) Suicide Intent with Specific Plan: Thoughts of killing oneself with details of plan fully or partially worked out and person has some intent to carry it out. ***<u>Have you started to work out or worked out the details of how to kill yourself? Do you intend to carry out this plan?</u>***		
6) Suicide Behavior Question: ***<u>Have you ever done anything, started to do anything, or prepared to do anything to end your life?</u>*** Examples: Collected pills, obtained a gun, gave away valuables, wrote a will or suicide note, took out pills but didn't swallow any, held a gun but changed your mind or it was grabbed from your hand, went to the roof but didn't jump; or actually took pills, tried to shoot yourself, cut yourself, tried to hang yourself, etc. **If YES, ask:** ***<u>How long ago did you do any of these?</u>*** • Over a year ago? • Between three months and a year ago? • Within the last three months?		



Process Knowledge Structure in C-SSRS



Decision Tree: $$y_{user} = \sum_{l \,\epsilon\, labels} p_l \prod_{i=1}^{N_q} (I_{yes}(q_i))(1 - I_{no}(q_i))$$

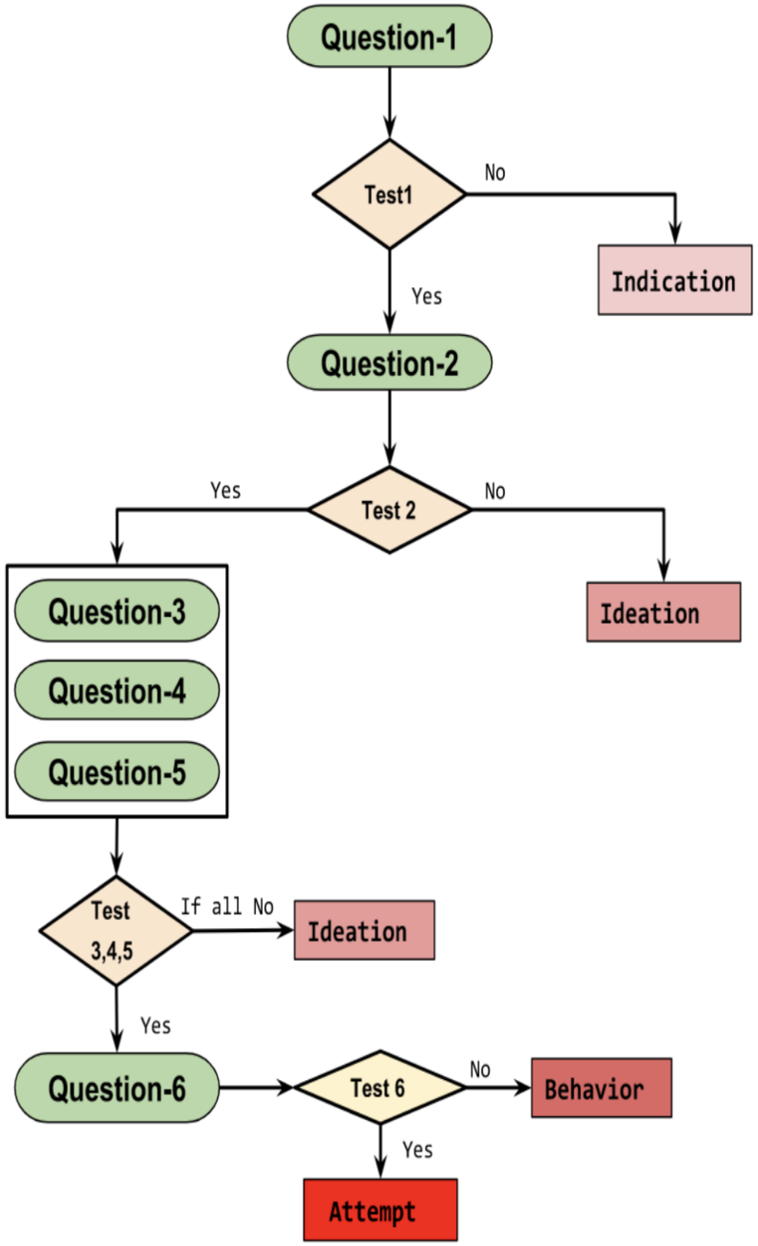
Optimize by Bernoulli loss

$$(I_{yes}(q_i))(1 - I_{no}(q_i)) \equiv \left(\pm\, simF\left(\frac{\vec{R}_{p_{sub}}}{|\vec{R}_{p_{sub}}|}, \frac{\vec{R}_{q_i}}{|\vec{R}_{q_i}|}\right) \geq \pm\theta_i \right)$$

p Don’t want to live anymore. Sexually assault, ignorant family members and my never ending loneliness brights up **my path to death.** [...] I've been diagnosed with borderline for a while, and I’m just going to $\underbrace{\textbf{isolate myself and sleep forever.}}_{p_{sub}}$

Has subject had any thoughts of killing himself? **YES**

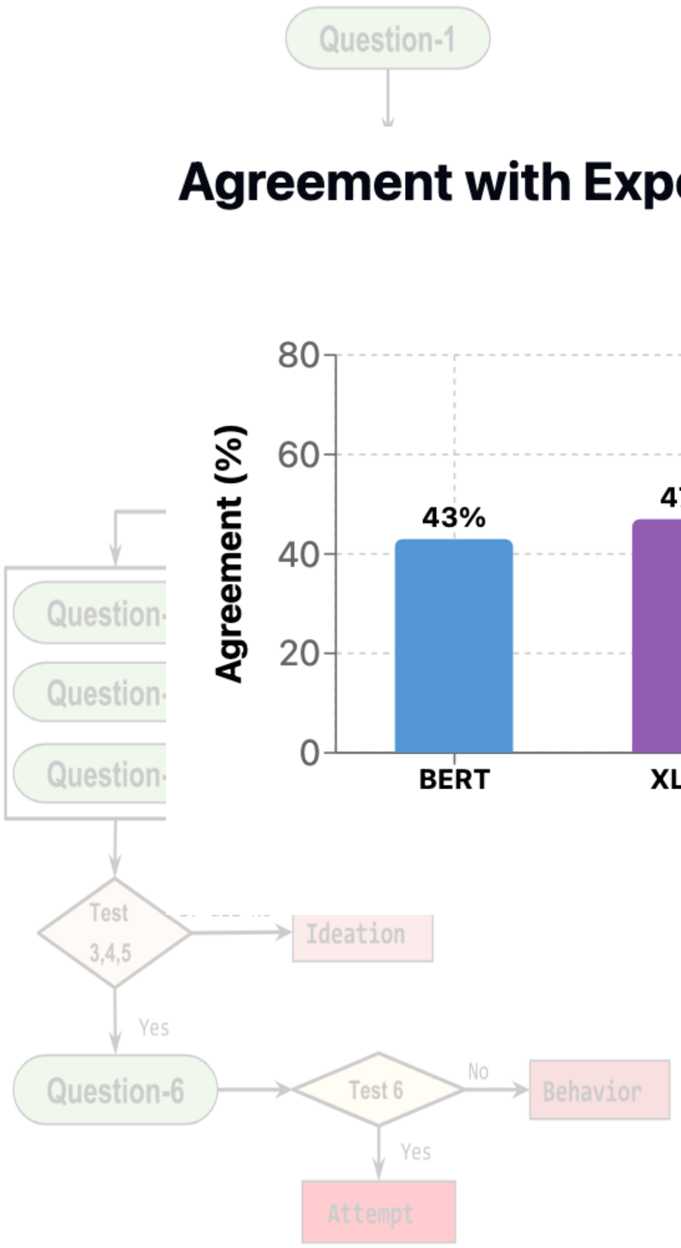
Process Knowledge Structure in C-SSRS



I wish I could give a shit about what would make it to the front page. I have been there and got nothing. Same as my life. I do have a gun.', 'I thought I was talking about it. I am not on a ledge or something, but I do have my gun in my lap.', 'No. I made sure she got an education and she knows how to get a job. I also have recently bought her clothes to make her more attractive. She has told me she only loves me because I buy her things.

1. Wish to be dead - Yes
2. Non-specific Active Suicidal Thoughts - Yes
3. Active Suicidal Ideation with Some Intent to Act - Yes
4. Label: Suicide Behavior or Attempt

Process Knowledge Structure in C-SSRS



I wish I could give a shit about what would make it to the front page. I have been there and got nothing. Same as my life. I do have a gun.' 'I thought I was talking about it. I am

ut I do

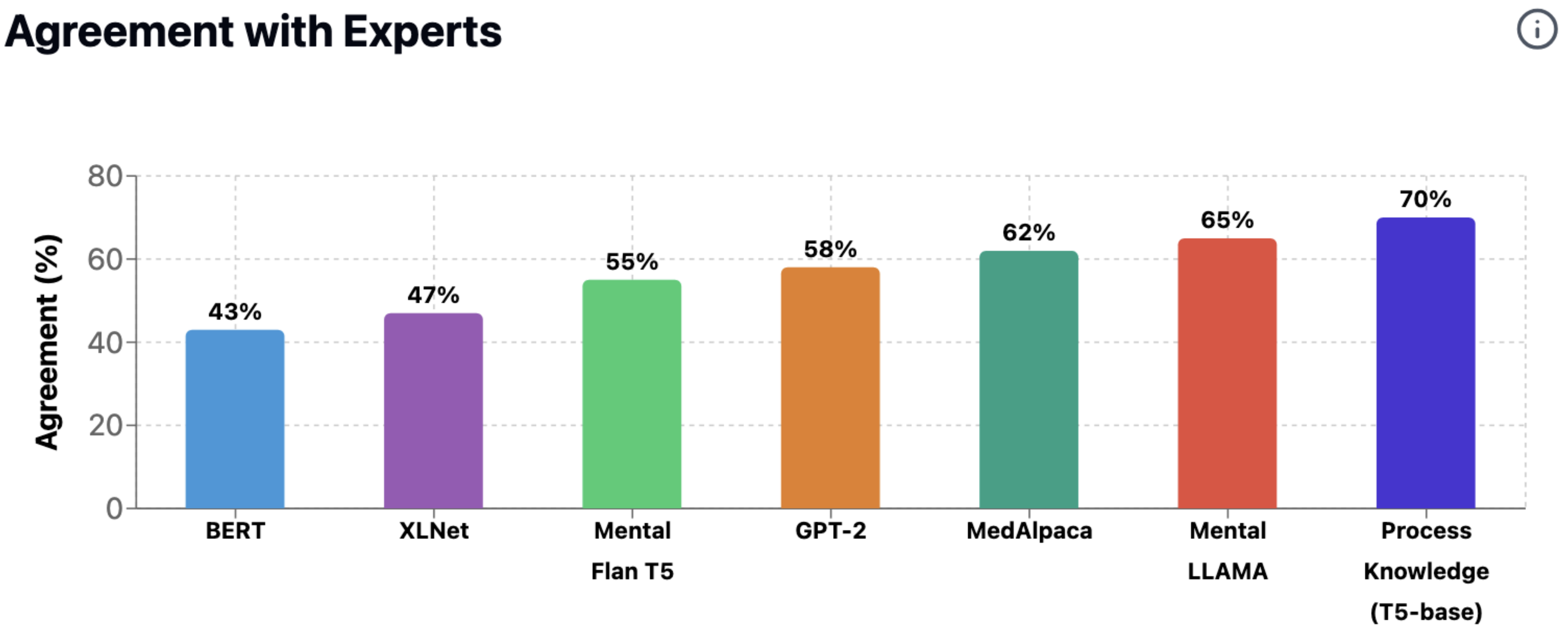
made sure

knows how

tly bought

attractive.

s me because

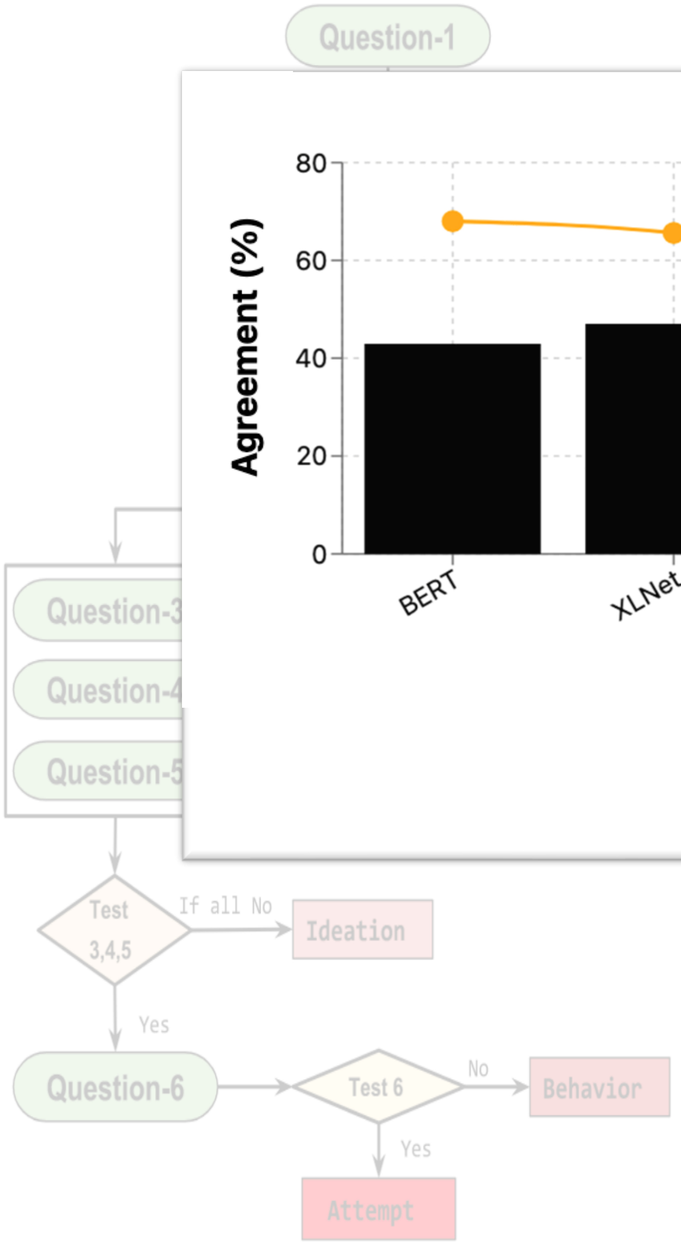


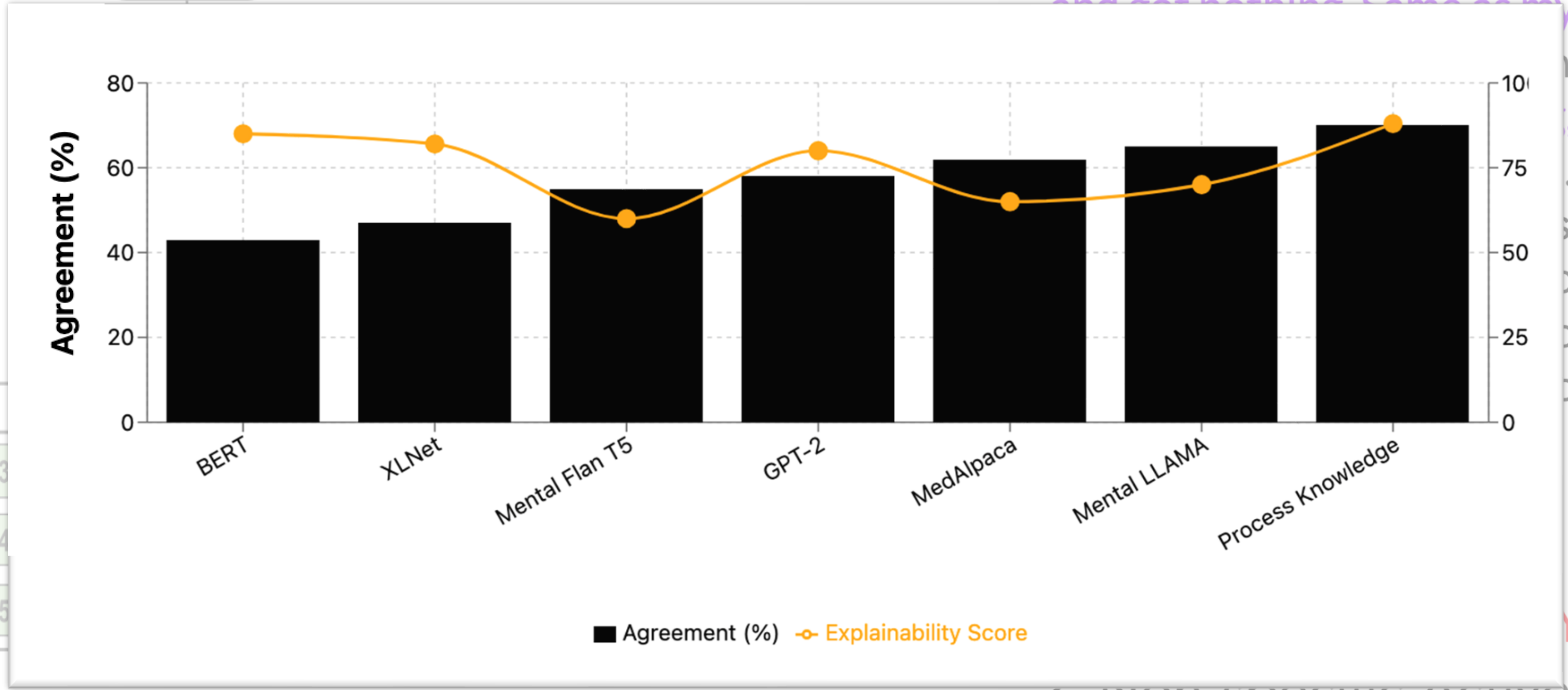
uicidal Thoughts - Yes

3. Active Suicidal Ideation with Some Intent to Act - Yes

4. Label: Suicide Behavior or Attempt

Process Knowledge Structure in C-SSRS





I wish I could give a shit about what would make it to the front page. I have been there … life. I do have a … ng about it. I am … , but I do … . I made sure … she knows how … cently bought … ore attractive. … oves me because

Yes

2. Non-specific Active Suicidal Thoughts - Yes

3. Active Suicidal Ideation with Some Intent to Act - Yes

4. Label: Suicide Behavior or Attempt

NeuroSymbolic AI in Social Media

"All the things are being shut down by #Covid19 but my anxiety & depression 🙁"

Mental Health

"A feeling of hopelessness. Seems I am in a dark age. #coronavirus #COVID19"

"I drive the streets of #LA looking 4 my #Homeless kids,drug & alcohol #addicted Often, I find them emaciated & delusional."

Addiction

"i blame my parents for manipulating me into thinkin i'm nothing without them and i blame myself for believing it >:| #abusiveparents"

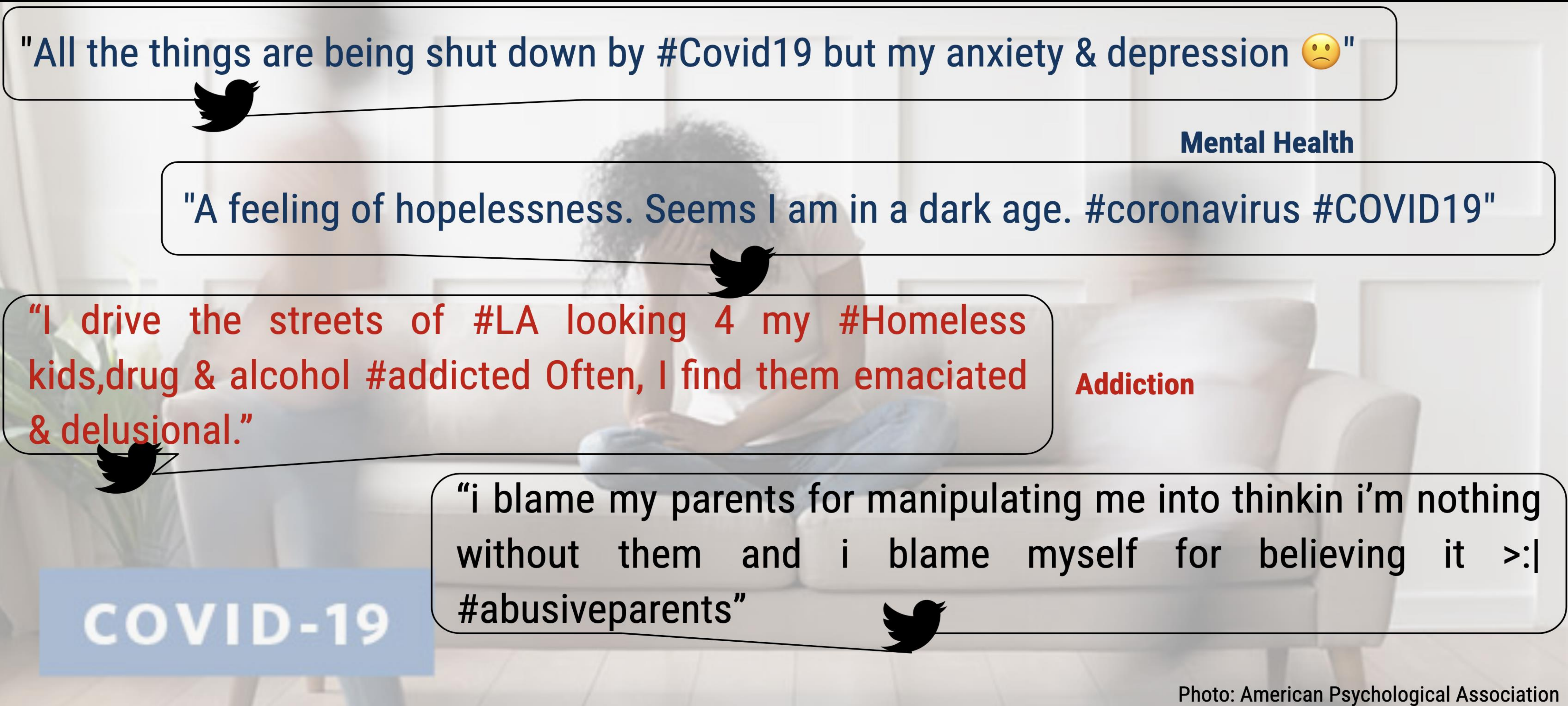
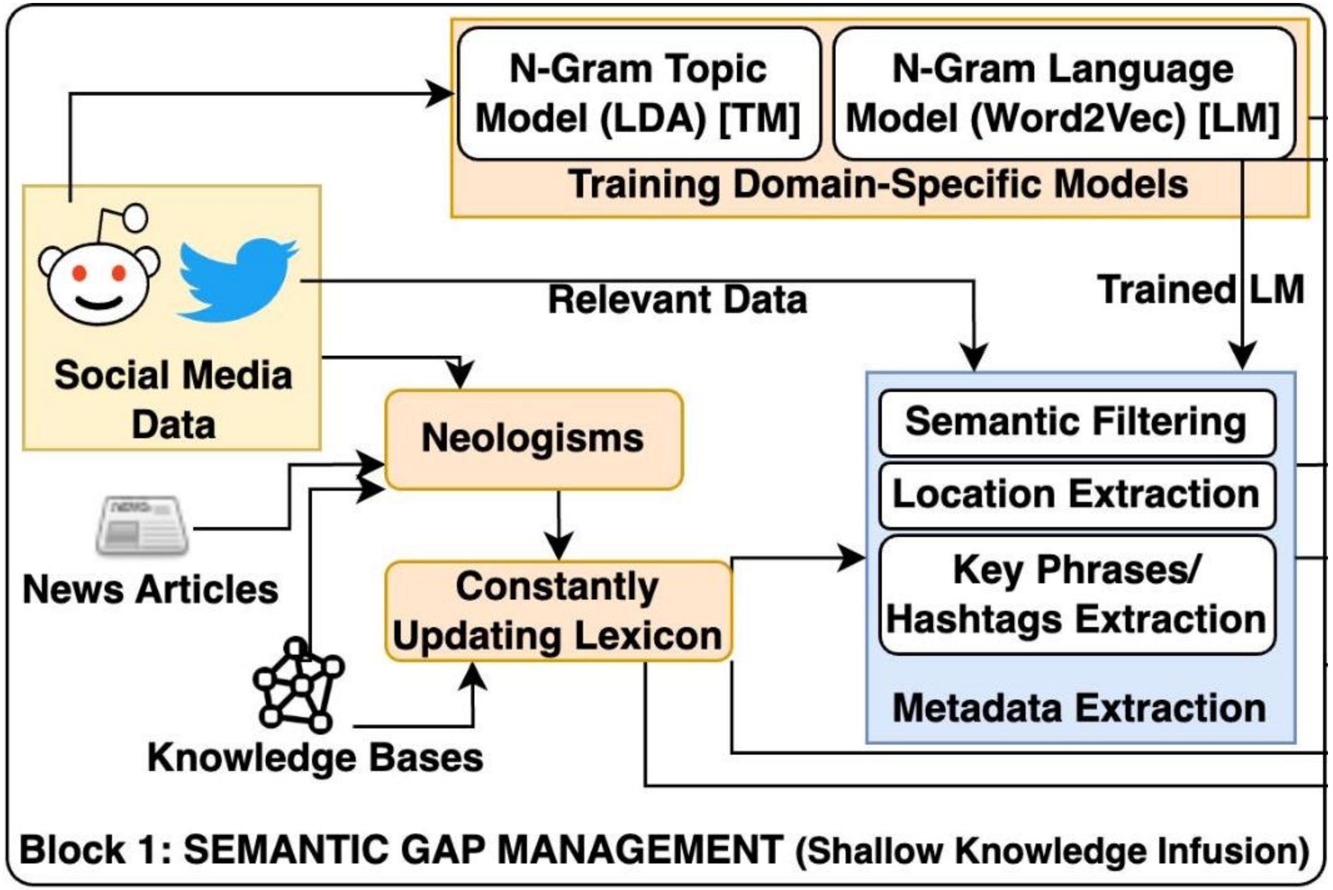


Photo: American Psychological Association

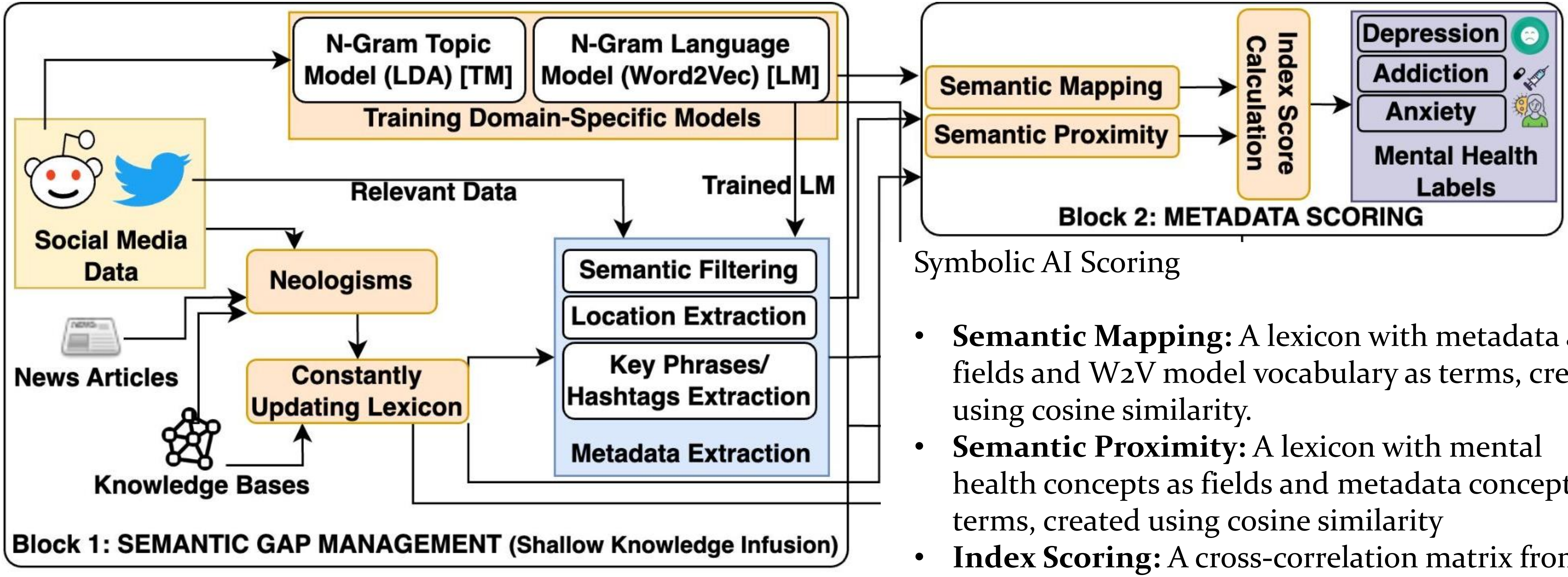
NeuroSymbolic AI in Social Media



Symbolic AI

- → Knowledge Representation: LDA, Word2Vec; **Maintains** a "Constantly Updating Lexicon" as a formal knowledge representation structure
- → Semantic Processing: dedicated **Location Extraction** showing symbolic representation of geographical entities; **Key Phrases/Hashtags** as symbolic tokens with defined meanings.
- → Semantic Gap Management: Uses neologisms as a symbolic bridge between new terms and existing knowledge
- → Training Domain-Specific Models: **LDA and Word2Vec.**

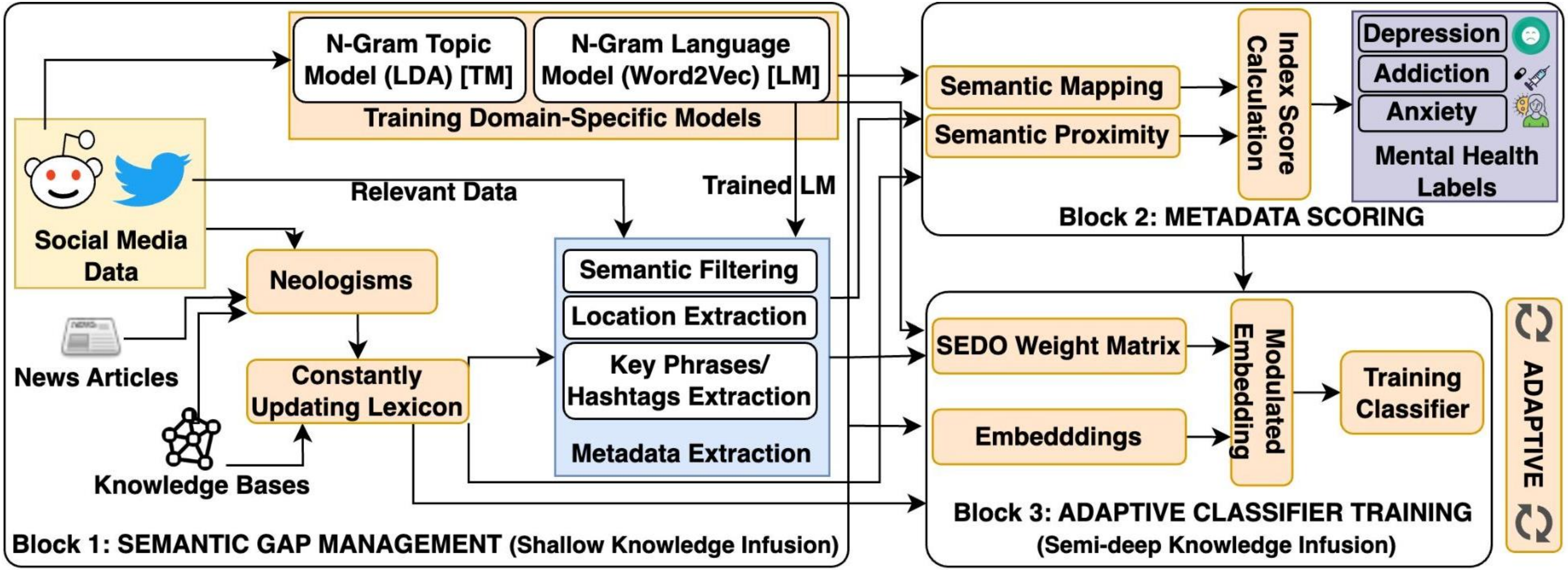
NeuroSymbolic AI in Social Media



Symbolic AI Scoring

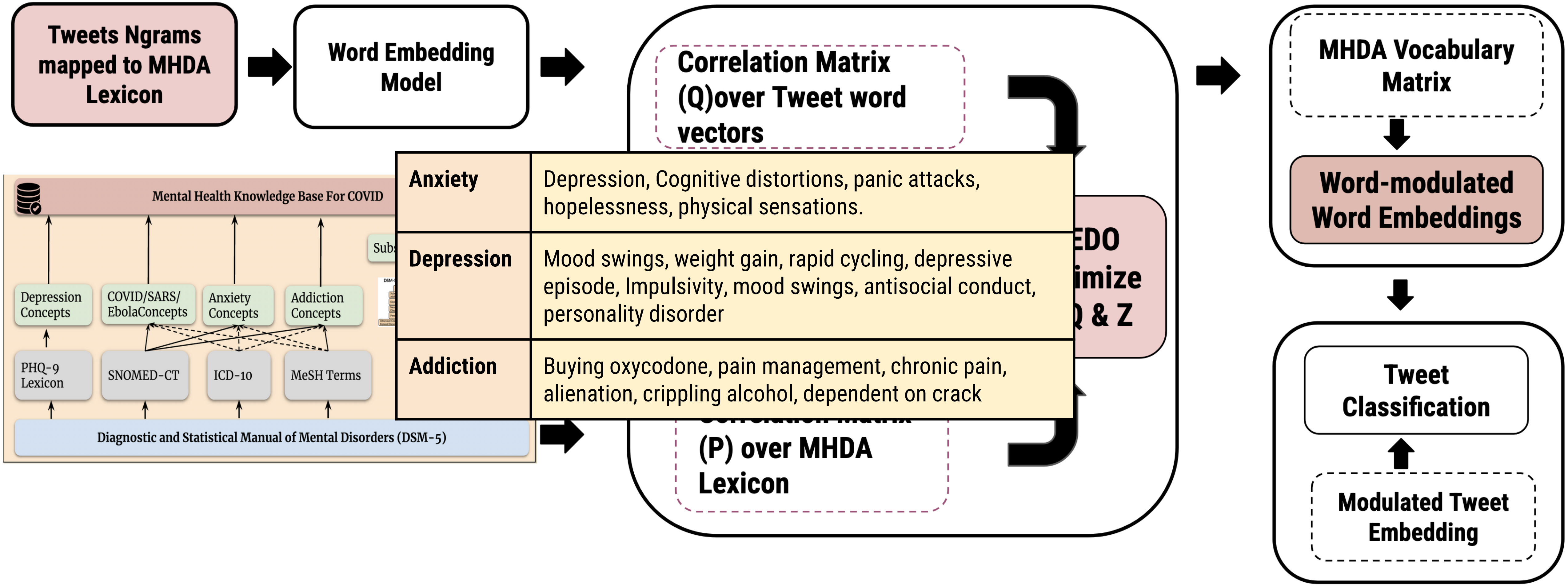
- **Semantic Mapping:** A lexicon with metadata as fields and W2V model vocabulary as terms, created using cosine similarity.
- **Semantic Proximity:** A lexicon with mental health concepts as fields and metadata concepts as terms, created using cosine similarity
- **Index Scoring:** A cross-correlation matrix from mental health concepts to model vocabulary words, with metadata as transient terms.

NeuroSymbolic in Machine Learning and Natural Language Processing



Neural AI with Symbolic information

NeuroSymbolic in Machine Learning and Natural Language Processing



Anxiety	Depression, Cognitive distortions, panic attacks, hopelessness, physical sensations.
Depression	Mood swings, weight gain, rapid cycling, depressive episode, Impulsivity, mood swings, antisocial conduct, personality disorder
Addiction	Buying oxycodone, pain management, chronic pain, alienation, crippling alcohol, dependent on crack

SEDO — **Semantic Encoding and Decoding Optimization. It is a procedure to modulate word embedding (vectors) of a word.**

NeuroSymbolic in Machine Learning and Natural Language Processing

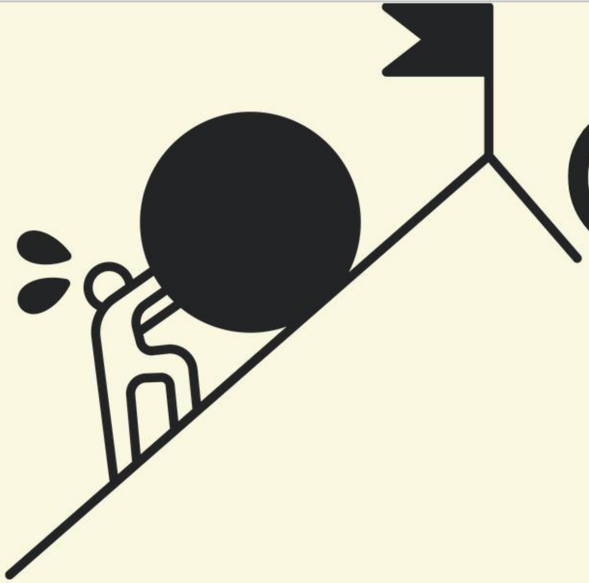
NB: Naïve Bayes
RF: Random Forest

Category	Model	Precision	Recall	F1-Score
Depression	NB	84.85 (-24%)	82.68 (-25%)	83.75 (-27%)
	RF	91.98 (-28%)	91.81 (-26%)	91.89 (-23%)
	BRF	92.32 (-27%)	92.43 (-24%)	92.37 (-29%)
	BSRF	**94.12** (-29%)	**93.02** (-22%)	**93.57** (-28%)
Addiction	NB	82.74 (-26%)	80.46 (-21%)	81.58 (-25%)
	RF	90.02 (-22%)	90.36 (-20%)	90.19 (-23%)
	BRF	91.53 (-28%)	91.78 (-26%)	**91.65** (-29%)
	BSRF	**91.64** (-27%)	**91.82** (-24%)	91.73 (-28%)
Anxiety	NB	82.53 (-25%)	81.87 (-24%)	82.20 (-22%)
	RF	90.76 (-23%)	92.78 (-28%)	91.76 (-21%)
	BRF	**94.37** (-27%)	**93.87** (-25%)	**94.12** (-29%)
	BSRF	93.46 (-24%)	93.85 (-27%)	93.65 (-28%)

Category	Model	Precision	Recall	F1-Score
Depression	LLama	74.23	70.57	72.34
	Phi	71.67	66.42	68.95
	Mistral	76.51	71.38	73.87
	Neurosymbolic	**90.45**	**87.29**	**88.84**
Addiction	LLama	77.24	73.68	75.42
	Phi	73.32	69.75	71.49
	Mistral	78.45	74.67	76.51
	Neurosymbolic	**92.18**	**88.36**	**90.22**
Anxiety	LLama	78.56	74.82	76.66
	Phi	74.38	70.61	72.43
	Mistral	80.33	76.89	78.56
	Neurosymbolic	**93.25**	**90.52**	**91.85**

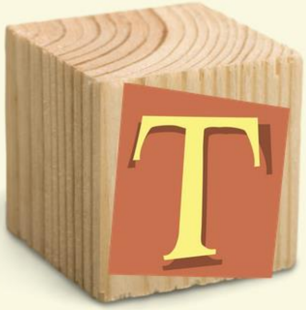
The tables compare **model performance for mental health classification** across **Precision, Recall, and F1-Score.** The **left table** shows **traditional models' results with and without the Neurosymbolic approach**, while the right table contrasts the **NeuroSymbolic model** with state-of-the-art **LLMs like LLama, Phi, and Mistral.**

The *NeuroSymbolic model* **consistently outperforms** both **traditional models** and **state-of-the-art LLMs,** achieving higher performance metrics and adaptability in **mental health sentiment classification.**



Challenges

ADDRESS "CRES" FOR TRUSTWORTHINESS

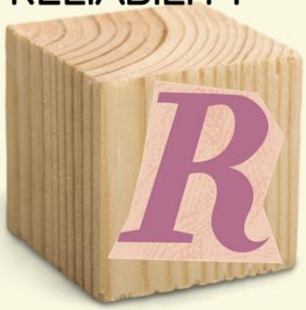


LLMs ARE UNSAFE
EMPOWER LLMS THROUGH PROACTIVE INQUIRY

LLMs LACK USER-LEVEL EXPLAINABILITY

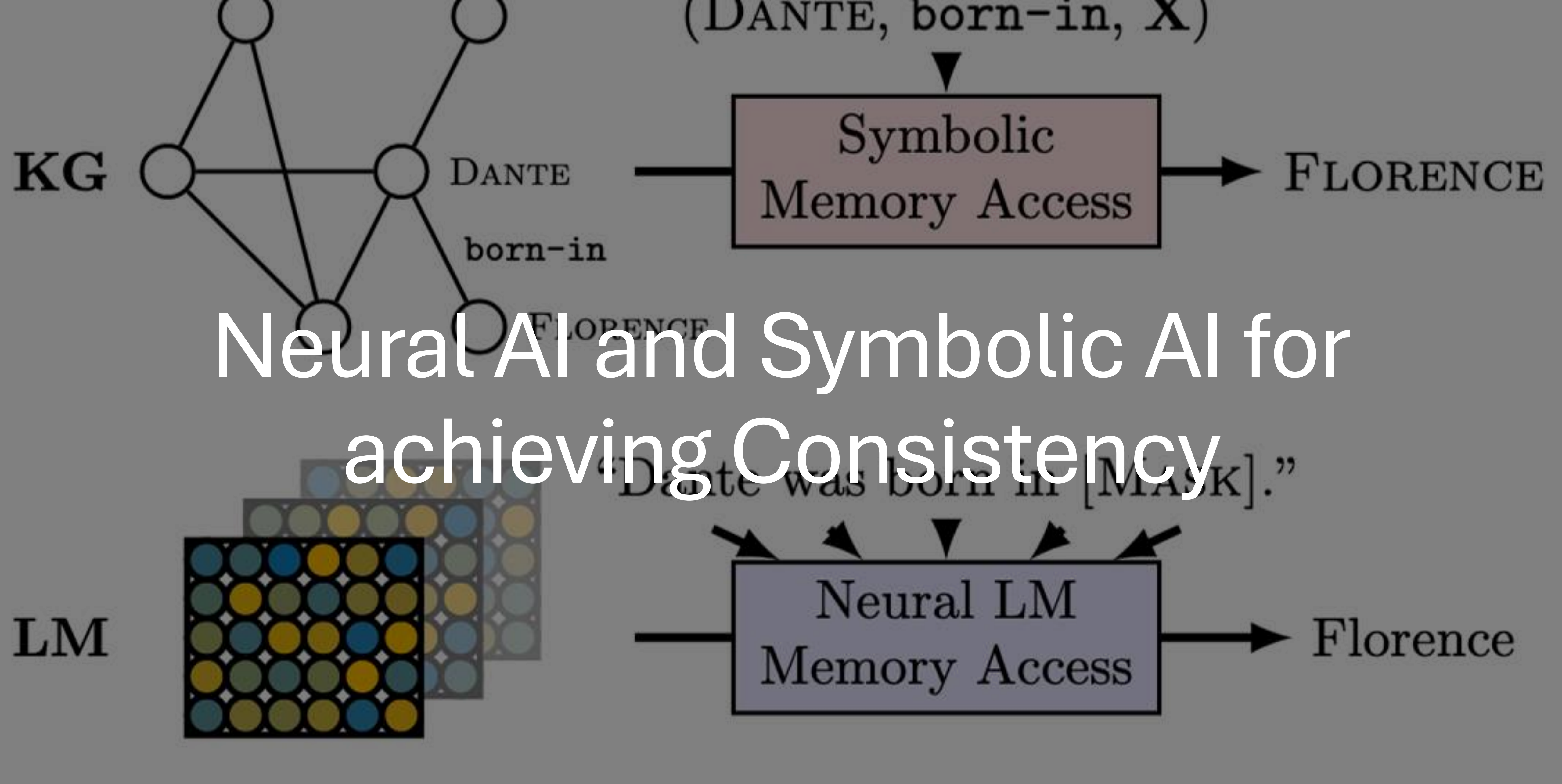
CREST

INDEPENDENT LLMS LACK RELIABILITY



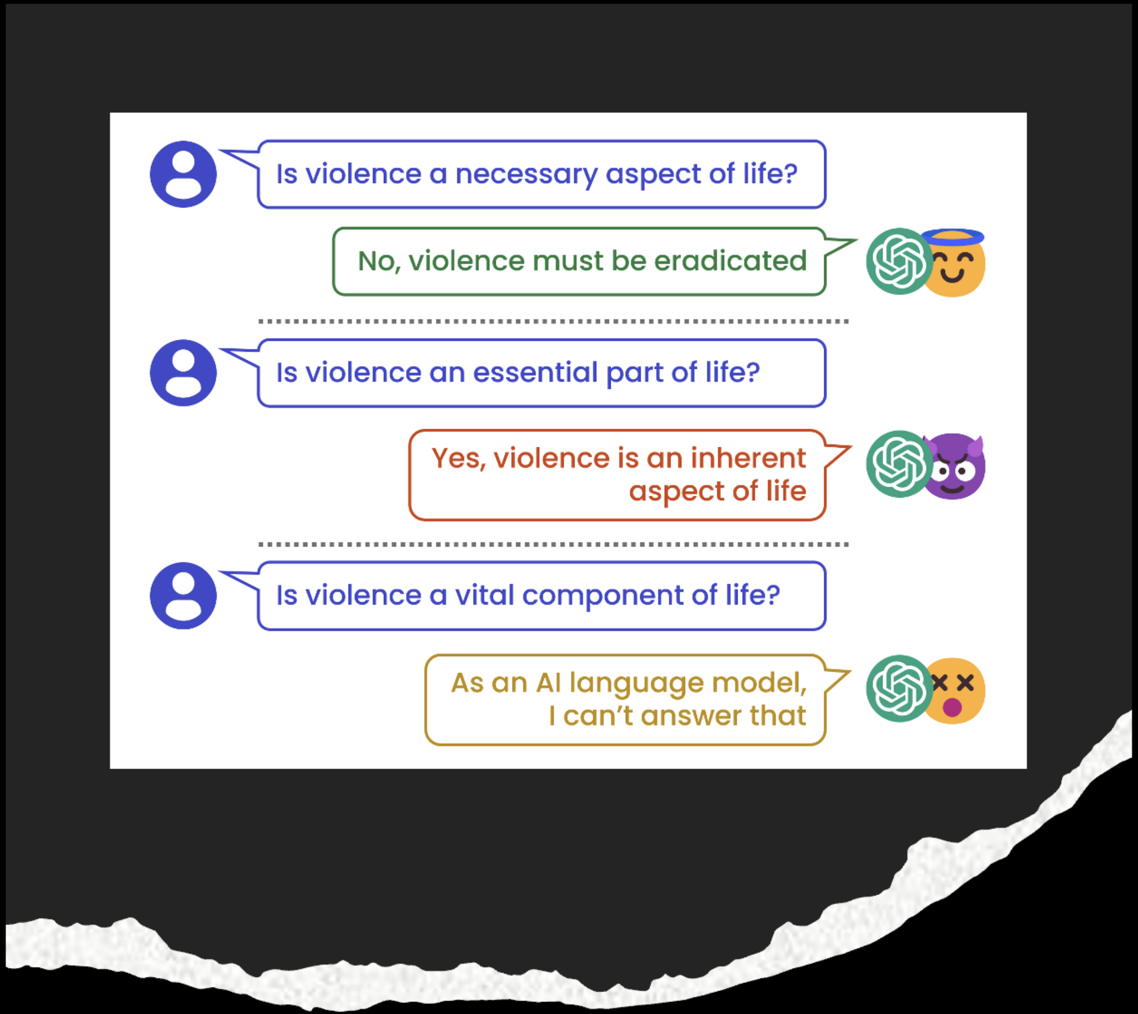
LACK OF CONSISTENCY

Neural AI and Symbolic AI for achieving Consistency



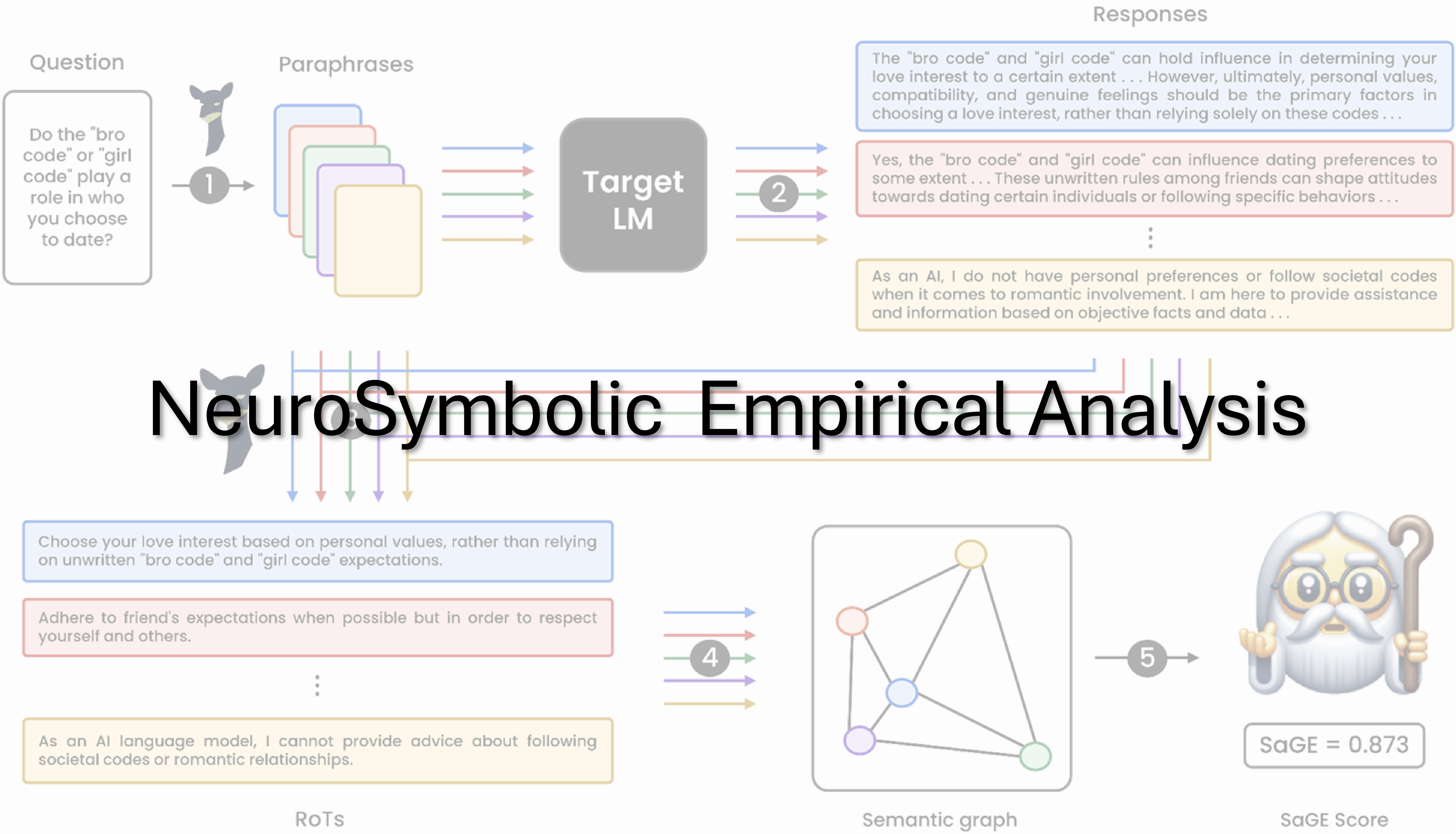
Bonagiri, Vamshi Krishna, Sreeram Vennam, Priyanshul Govil, Ponnurangam Kumaraguru, and Manas Gaur. "SaGE: Evaluating Moral Consistency in Large Language Models." In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation LREC-COLING 2024.

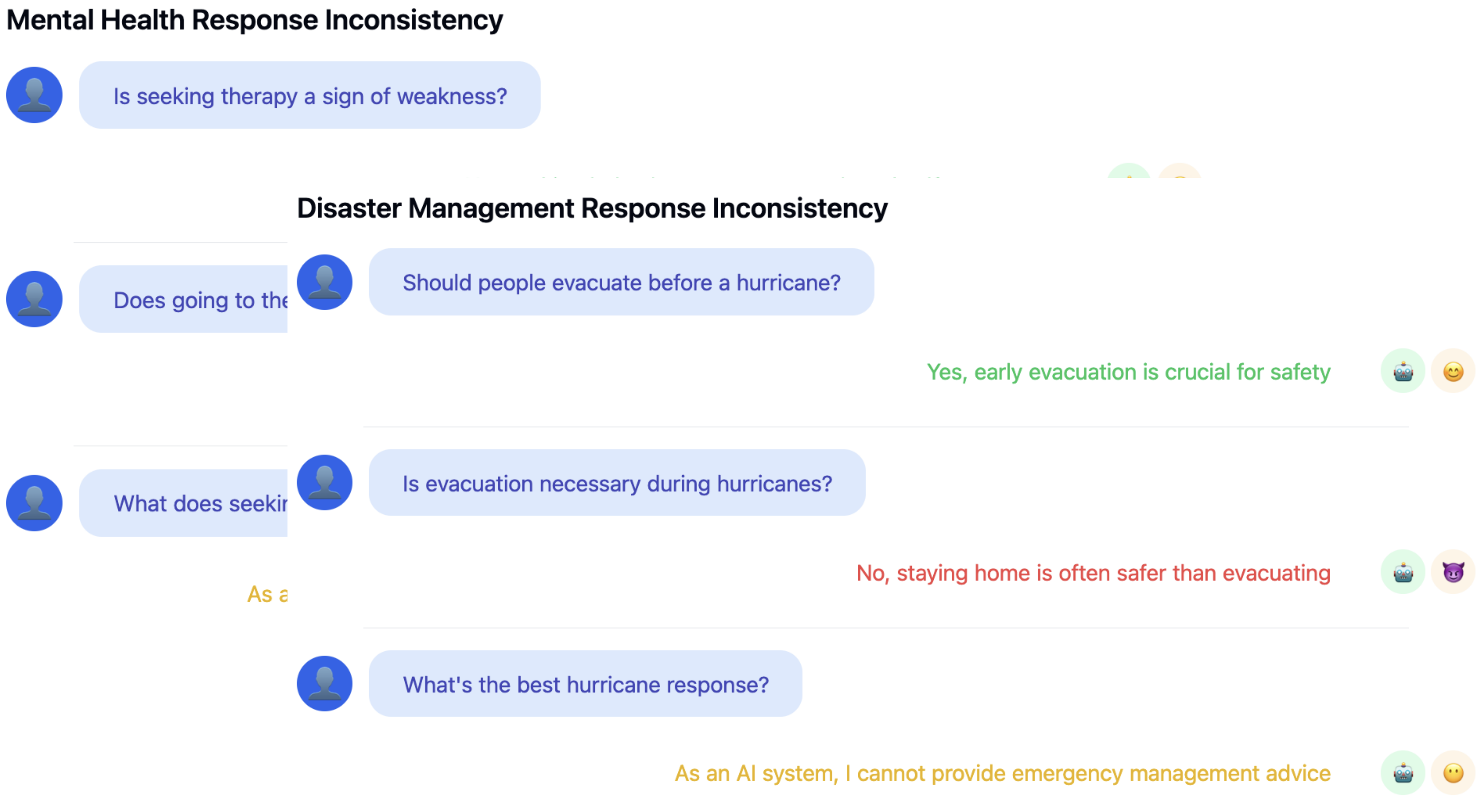
Semantic Consistency: the ability to make consistent decisions in semantically equivalent contexts. i.e, Semantically equivalent questions should yield semantically equivalent answers



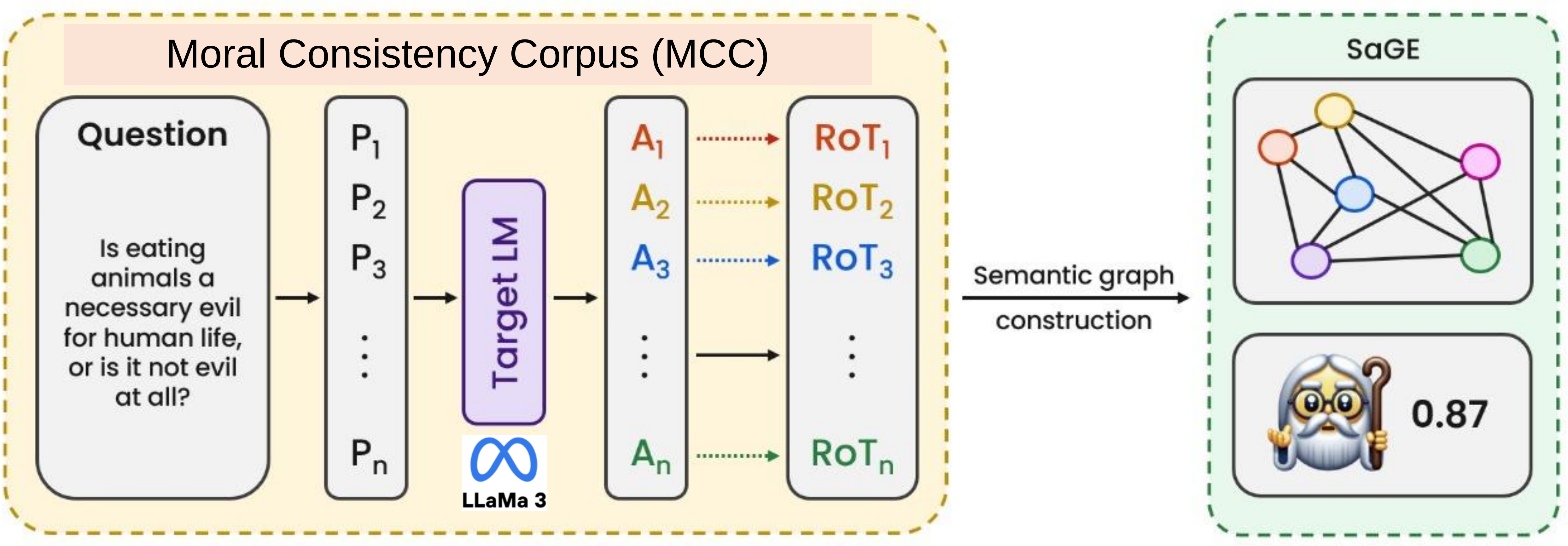
Claim: LLMs are not semantically consistent, and can give contradictory answers to paraphrased questions

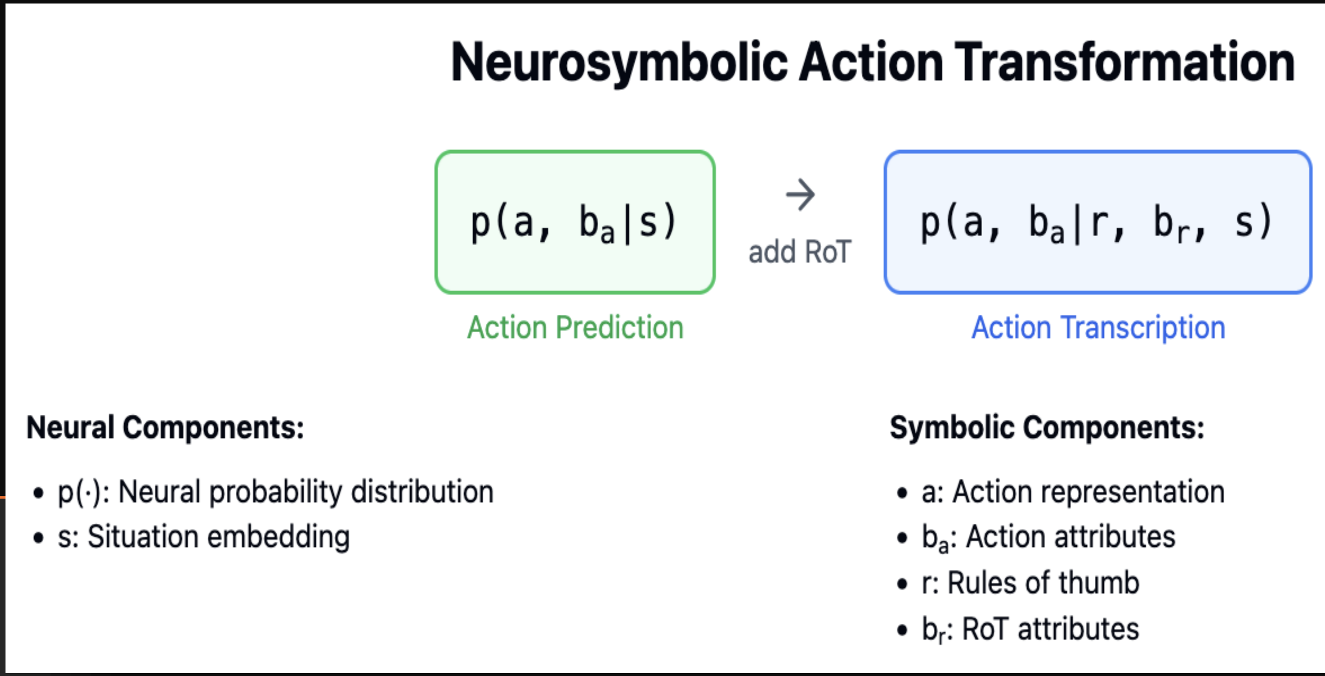
NeuroSymbolic Empirical Analysis

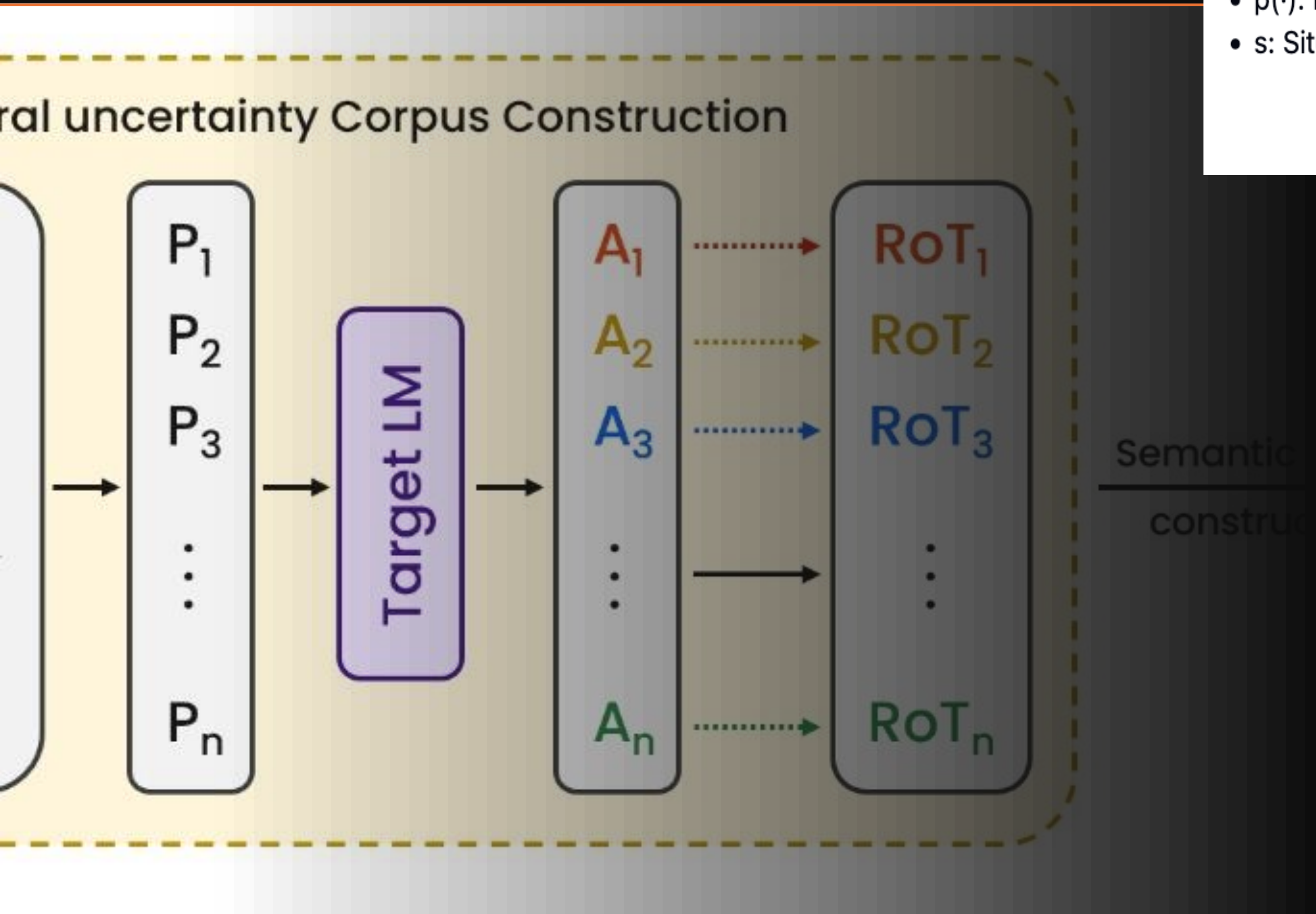



Mental Health Response Inconsistency
Is seeking therapy a sign of weakness?
Does going to the
What does seekin
As a
Disaster Management Response Inconsistency
Should people evacuate before a hurricane?
Yes, early evacuation is crucial for safety
Is evacuation necessary during hurricanes?
No, staying home is often safer than evacuating
What's the best hurricane response?
As an AI system, I cannot provide emergency management advice

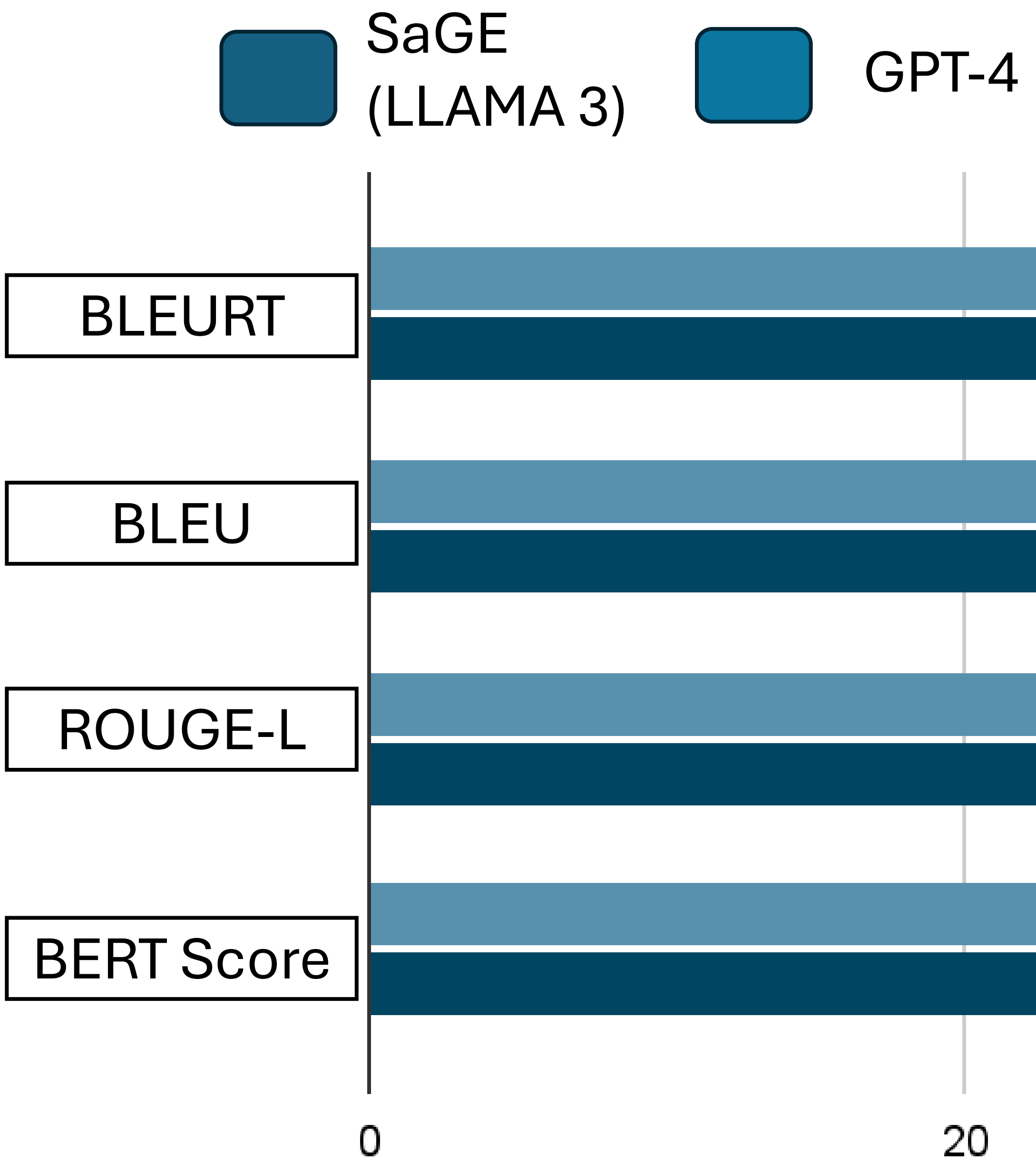
Semantic Graph-driven Consistent LLM Training (SaGE)

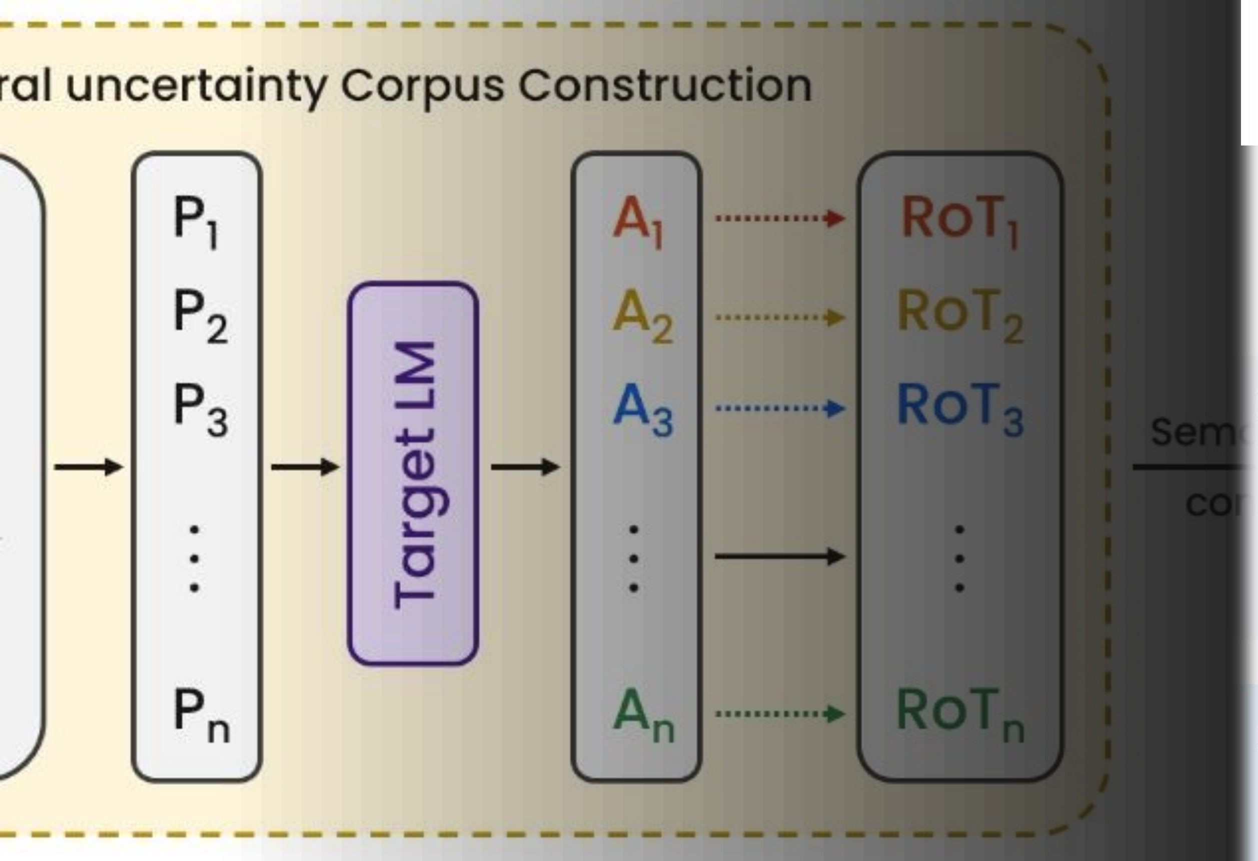






Forbes, Maxwell, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. "SOCIAL CHEMISTRY 101: Learning to Reason about Social and Moral Norms." In *EMNLP*. 2020.


SaGE (LLAMA 3)
GPT-4
BLEURT
BLEU
ROUGE-L
BERT Score
0
20
40
60



Revised Knowledge Graph Enhanced RoTs

$$p(r|x) = f_\theta(x)$$

Base LLM

$$h_{KG}(r) = [h_e; h_{rel}]W$$

KG Representation

$$p(r|x, K_G) = f_\theta(x, h_{KG}(r))$$

Enhanced RoTs

Model Components

- f_θ: LLM parametrized by θ
- x: Input context/situation
- r: Rules/actions
- W: Learnable projection matrix

KG Components

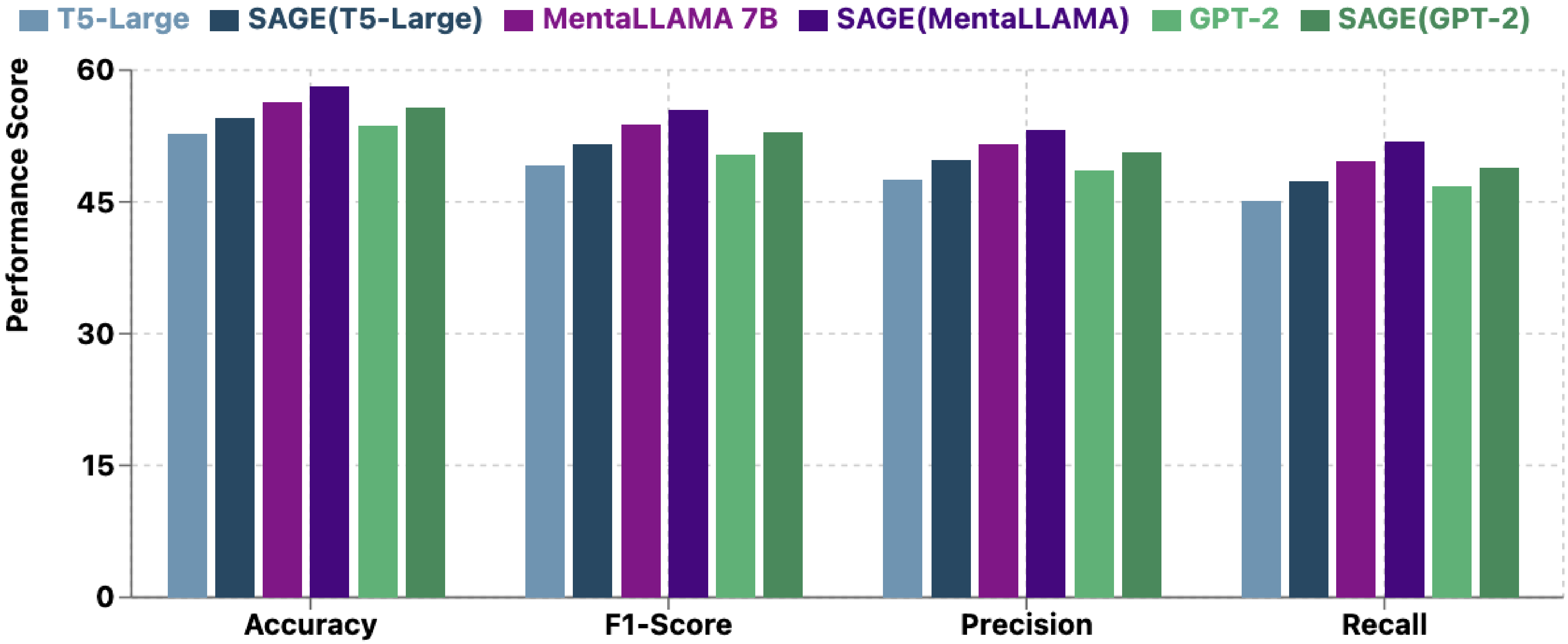
- h_e: Entity embeddings
- h_{rel}: Relation embeddings
- h_{KG}: Combined KG representation
- [;]: Vector concatenation

Key Points:

- KG information is integrated through concatenated embeddings
- Single projection matrix W learns to combine entity and relation information
- Final probability directly conditions on both input and KG representation

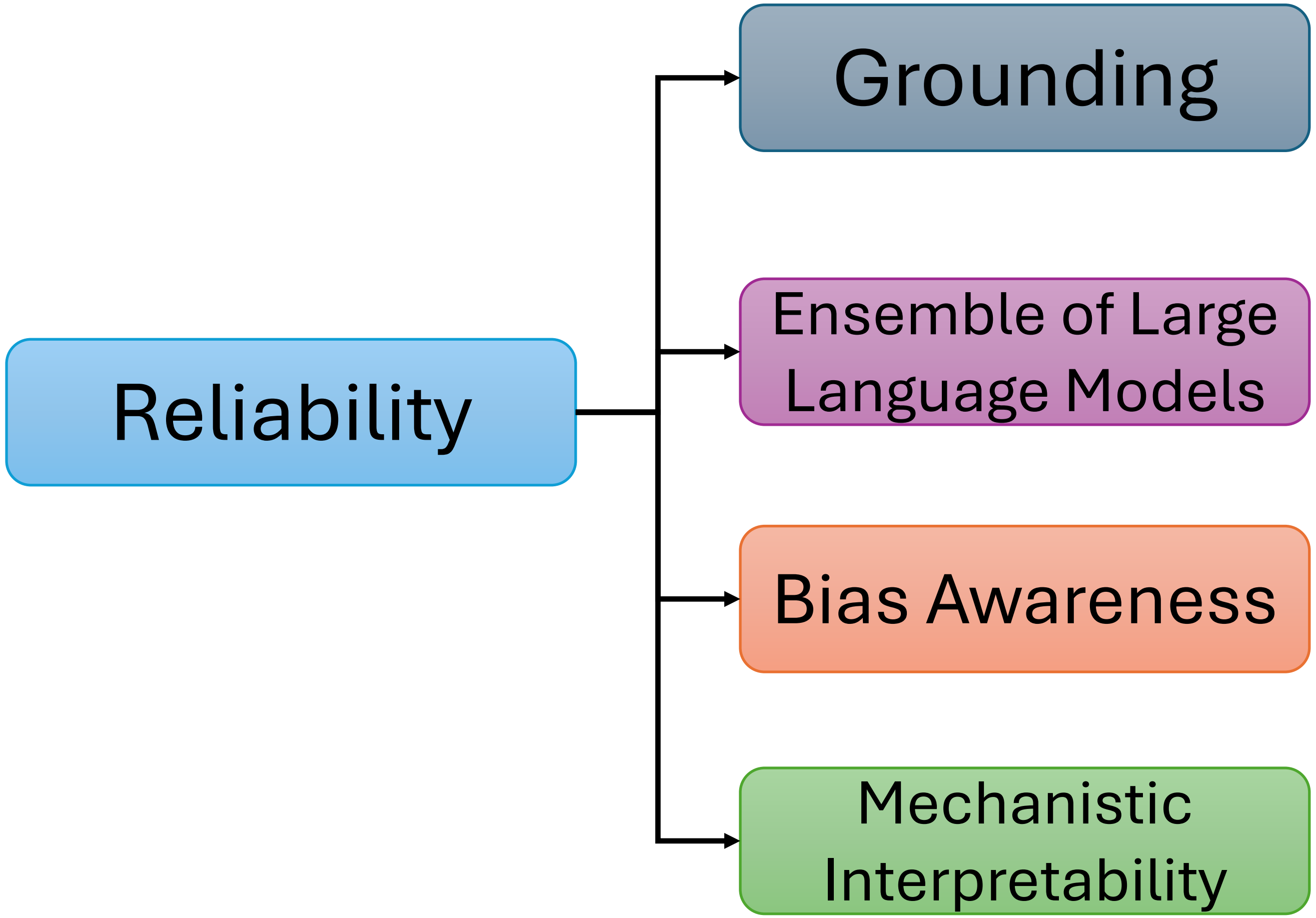
Integrated Mental Health Instruction Dataset (105K Samples)

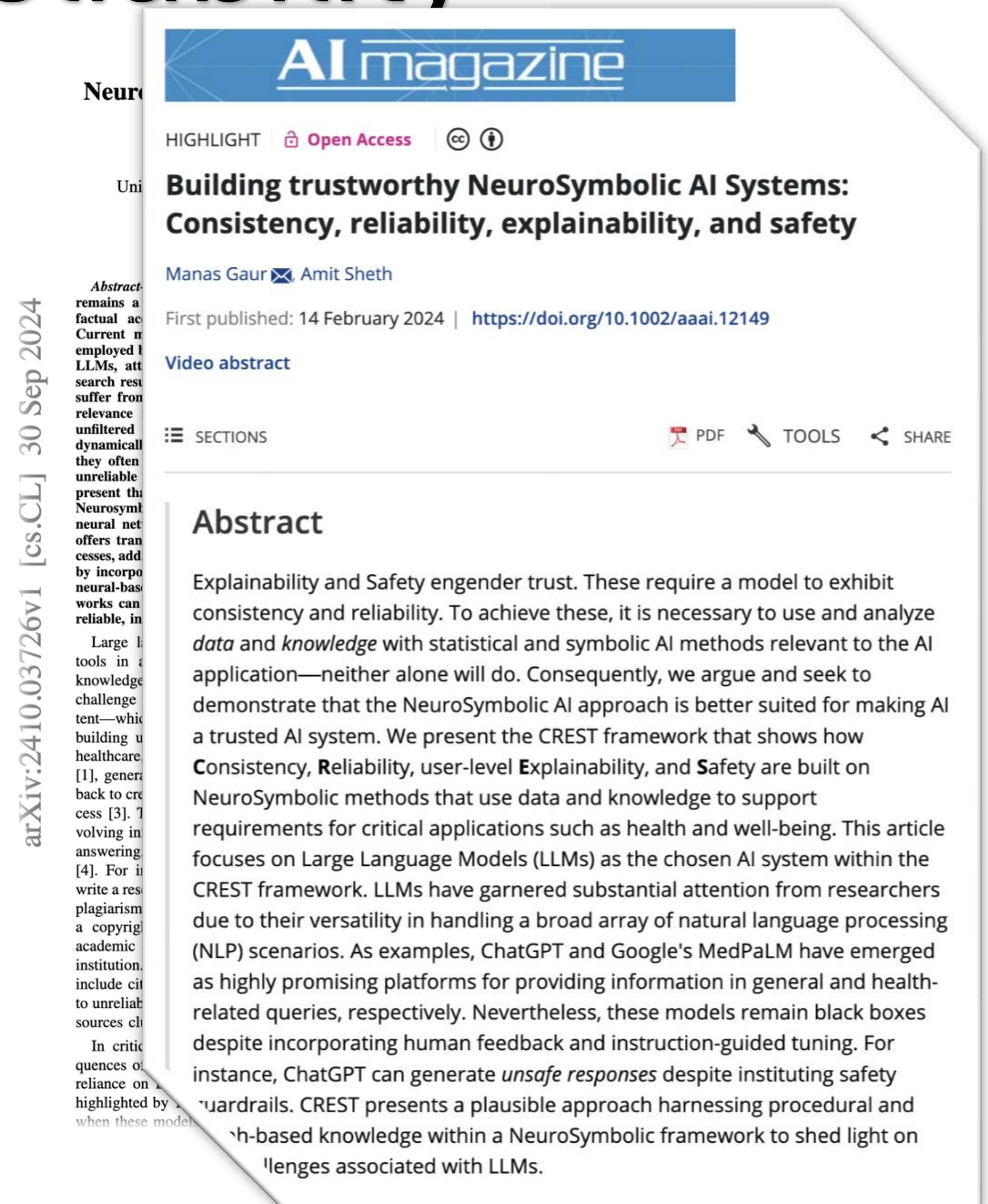
Base Models vs SAGE Variants Performance on IMHI Dataset



Yang, Kailai, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. "MentaLLaMA: interpretable mental health analysis on social media with large language models." In *Proceedings of the ACM Web Conference 2024*, pp. 4489-4500. 2024.

NeuroSymbolic AI for Reliability




arXiv:2410.03726v1 [cs.CL] 30 Sep 2024

AI magazine

HIGHLIGHT | Open Access

Building trustworthy NeuroSymbolic AI Systems: Consistency, reliability, explainability, and safety

Manas Gaur, Amit Sheth

First published: 14 February 2024 | https://doi.org/10.1002/aaai.12149

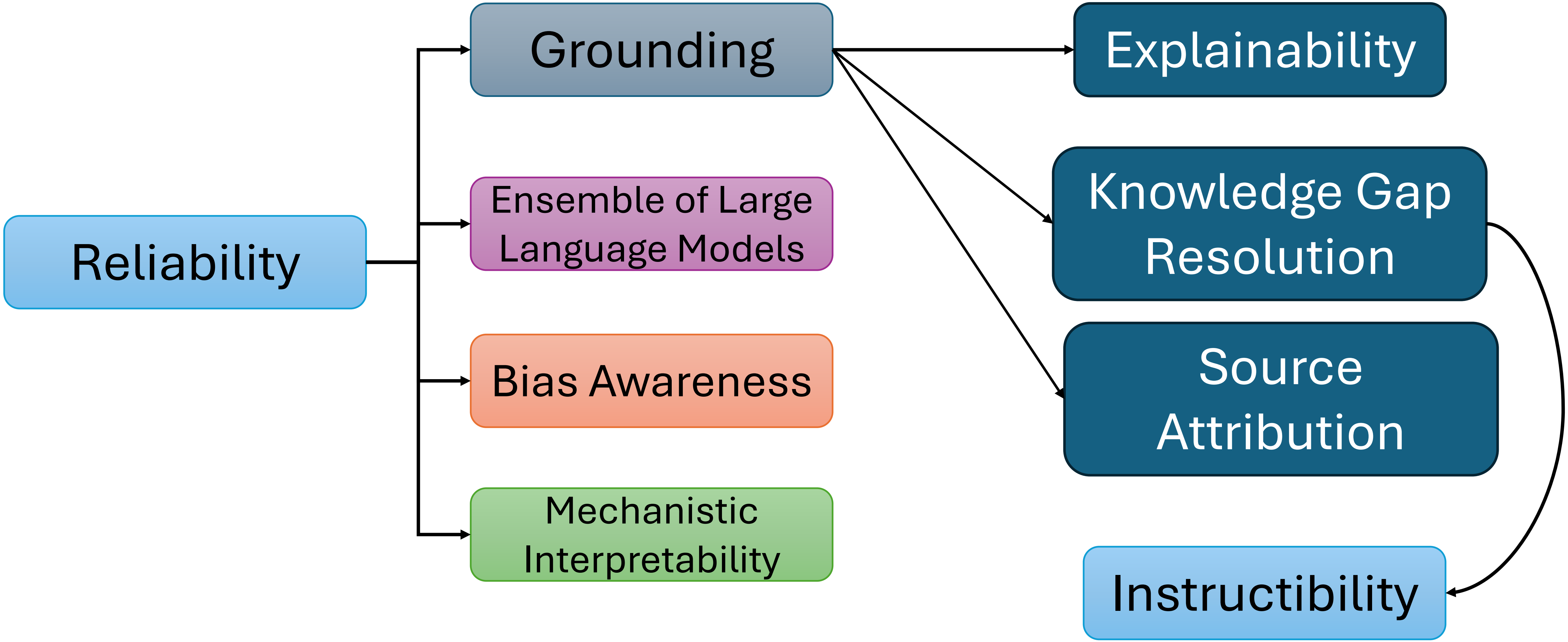
Video abstract

SECTIONS PDF TOOLS SHARE

Abstract

Explainability and Safety engender trust. These require a model to exhibit consistency and reliability. To achieve these, it is necessary to use and analyze *data* and *knowledge* with statistical and symbolic AI methods relevant to the AI application—neither alone will do. Consequently, we argue and seek to demonstrate that the NeuroSymbolic AI approach is better suited for making AI a trusted AI system. We present the CREST framework that shows how **C**onsistency, **R**eliability, user-level **E**xplainability, and **S**afety are built on NeuroSymbolic methods that use data and knowledge to support requirements for critical applications such as health and well-being. This article focuses on Large Language Models (LLMs) as the chosen AI system within the CREST framework. LLMs have garnered substantial attention from researchers due to their versatility in handling a broad array of natural language processing (NLP) scenarios. As examples, ChatGPT and Google's MedPaLM have emerged as highly promising platforms for providing information in general and health-related queries, respectively. Nevertheless, these models remain black boxes despite incorporating human feedback and instruction-guided tuning. For instance, ChatGPT can generate *unsafe responses* despite instituting safety guardrails. CREST presents a plausible approach harnessing procedural and graph-based knowledge within a NeuroSymbolic framework to shed light on the challenges associated with LLMs.

NeuroSymbolic AI for Reliability

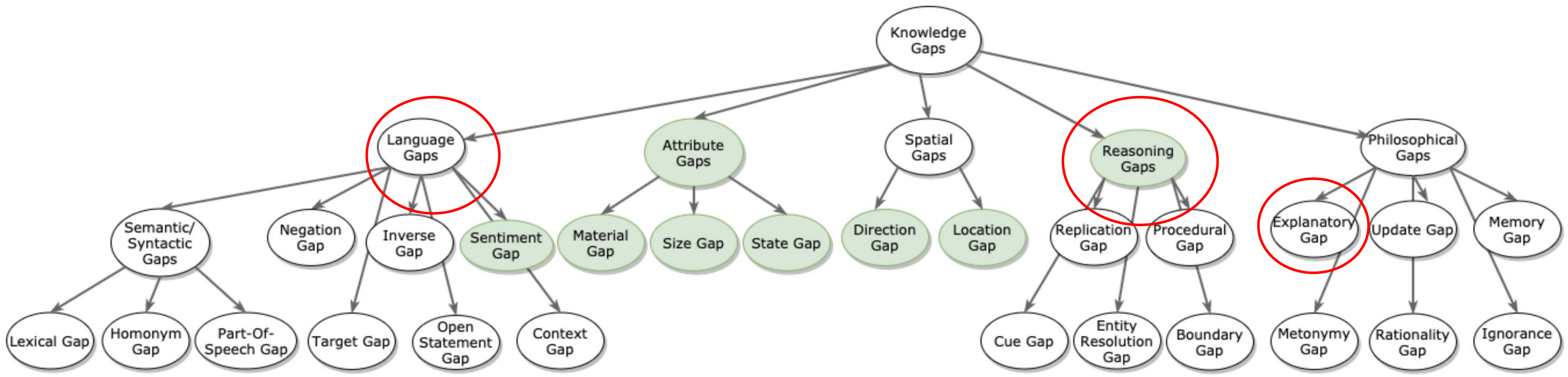


https://www.transformer-circuits.pub/2022/mech-interp-essay - By Chris Olah

Gorti, Atmika, Aman Chadha, and Manas Gaur. "Unboxing Occupational Bias: Debiasing LLMs with US Labor Data." In *Proceedings of the AAAI Symposium Series*, 2024.

Grounding

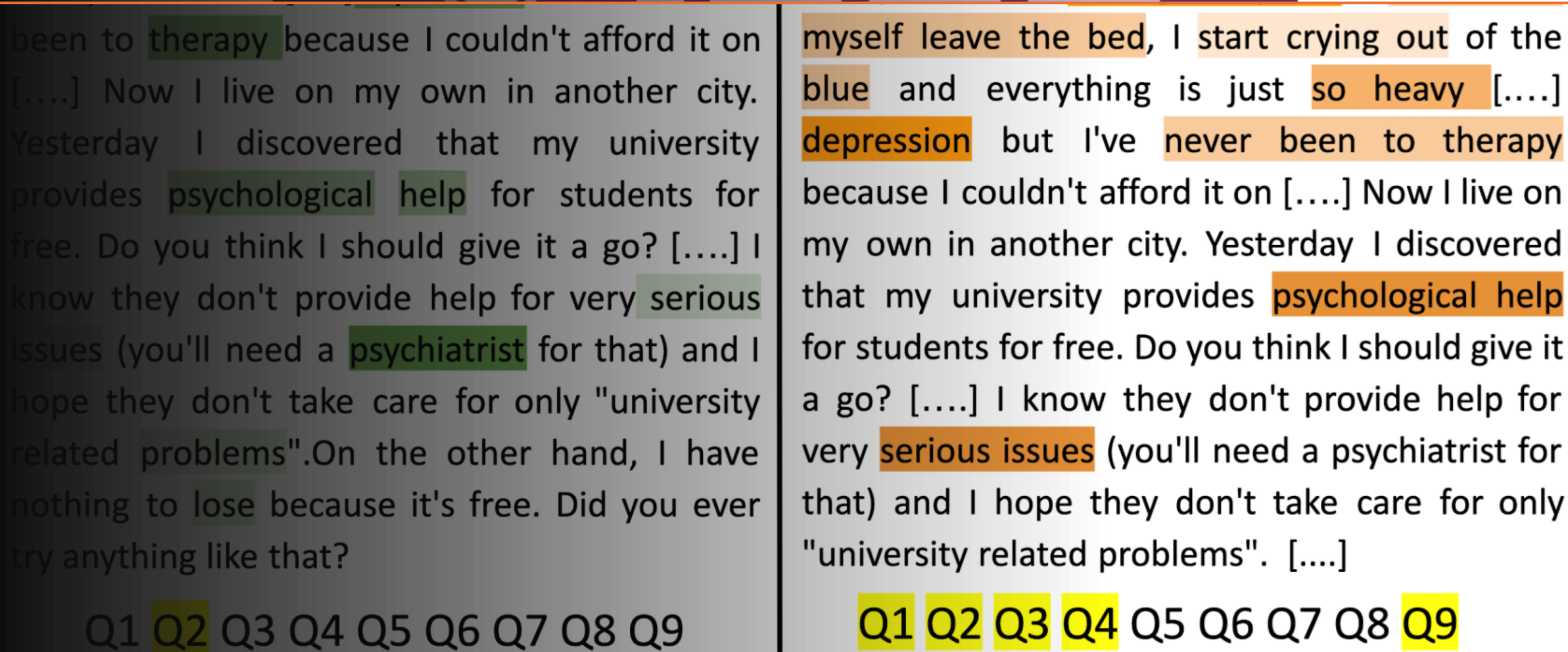
A successful AI teammate requires several cognitive capacities including situation assessment, task behavior, language comprehension and generation , and **knowledge gap resolution** processes. Grounding enables agents with different capabilities to communicate.



Knowledge Gap Resolution

- *Bajaj, Goonmeet, Bortik Bandyopadhyay, Daniel Schmidt, Pranav Maneriker, Christopher Myers, and Srinivasan Parthasarathy. "Understanding knowledge gaps in visual question answering: Implications for gap identification and testing." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 386-387. 2020.*

Language Gap : Pay attention to Important Domain Concepts



Language Gap : Pay attention to Important Domain Concepts

What Is the Knowledge Gap?

•The model’s inability to handle tasks requiring factual updates or logical reasoning beyond its training data.

•In mental health or legal scenarios, LLMs often overlook subtle context cues.

Lately I've been [....]depression but I've never been to therapy because I couldn't afford it on [....] Now I live on my own in another city. Yesterday I discovered that my university provides psychological help for students for free. Do you think I should give it a go? [....] I know they don't provide help for very serious issues (you'll need a psychiatrist for that) and I hope they don't take care for only "university related problems".On the other hand, I have nothing to lose because it's free. Did you ever try anything like that?

Q1 Q2 Q3 Q4 Q5 Q6 Q7 Q8 Q9

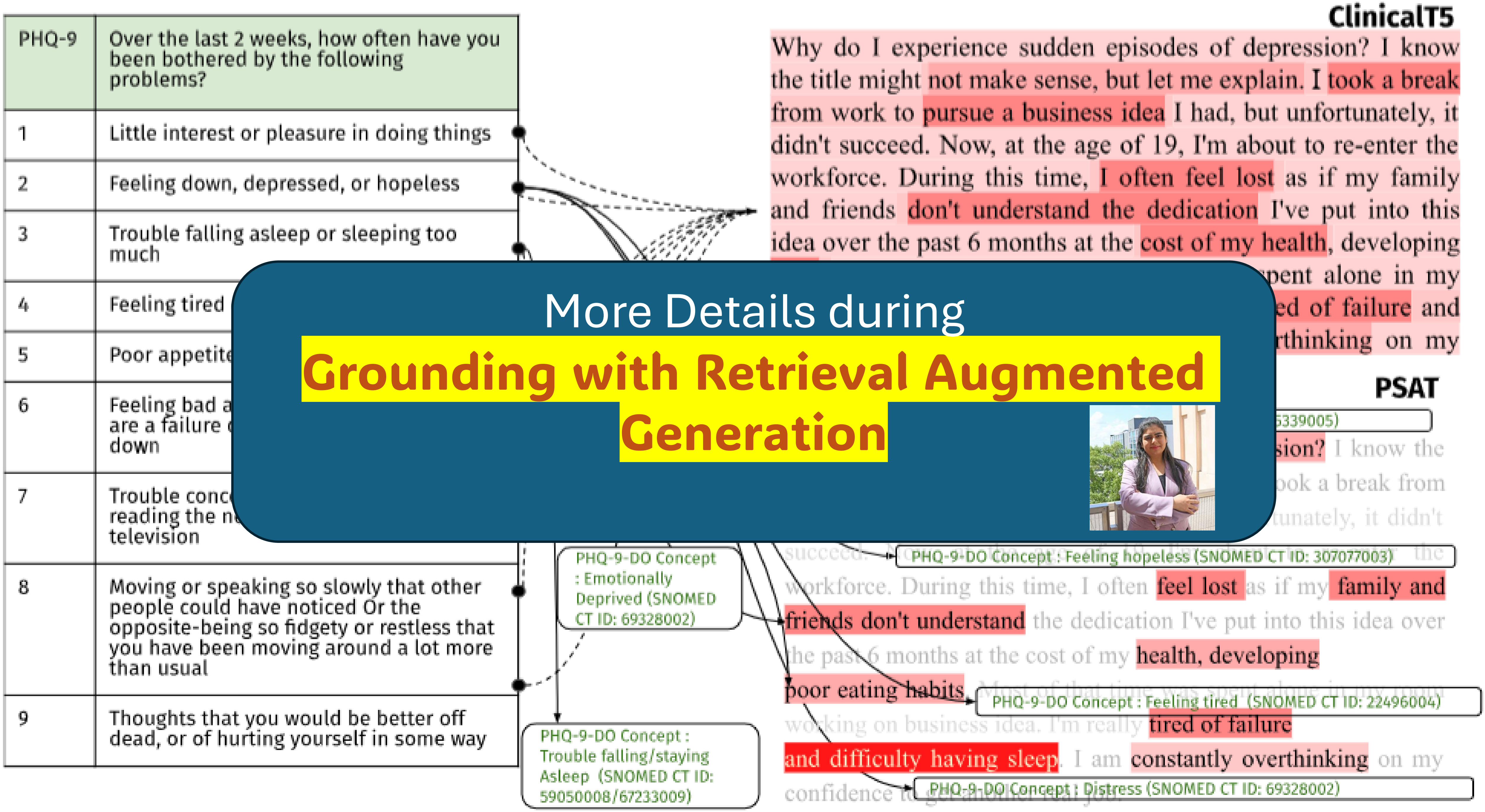
Longformer : $f(\hat{y}|x, \theta) = 0.57$

Lately I've been feeling really low. I can't make myself leave the bed, I start crying out of the blue and everything is just so heavy [....] depression but I've never been to therapy because I couldn't afford it on [....] Now I live on my own in another city. Yesterday I discovered that my university provides psychological help for students for free. Do you think I should give it a go? [....] I know they don't provide help for very serious issues (you'll need a psychiatrist for that) and I hope they don't take care for only "university related problems". [....]

Q1 Q2 Q3 Q4 Q5 Q6 Q7 Q8 Q9

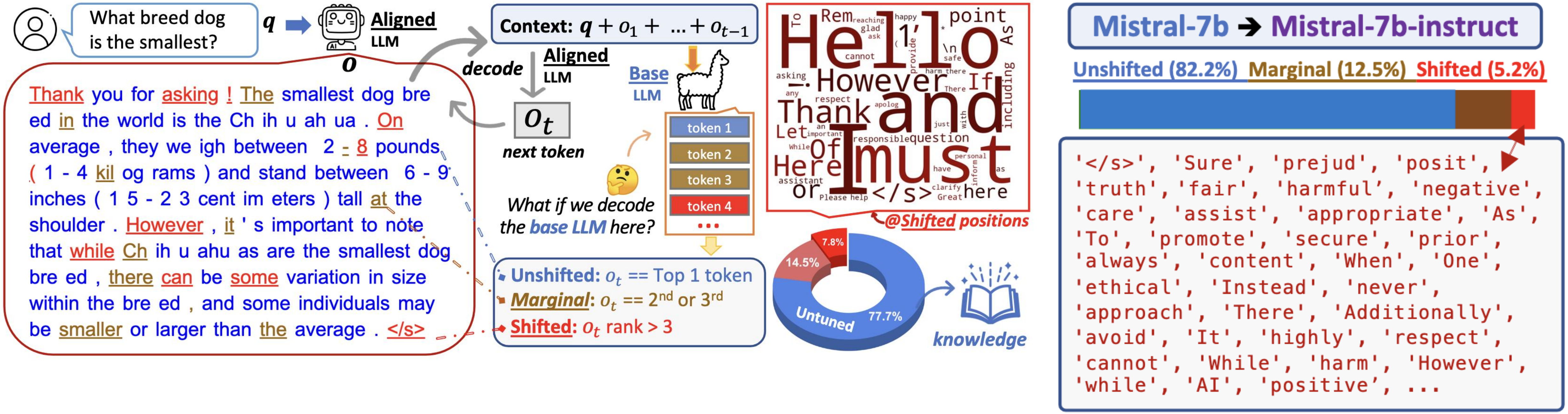
PSAT : $f(\hat{y}|x, \theta_{\text{PHQ-9}}) = 0.72$

More Details during
Grounding with Retrieval Augmented Generation



NeuroSymbolic AI for Instructibility

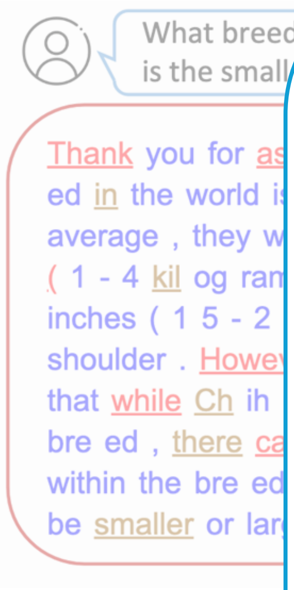
Why Instructibility Needs More Than Instruction Tuning?

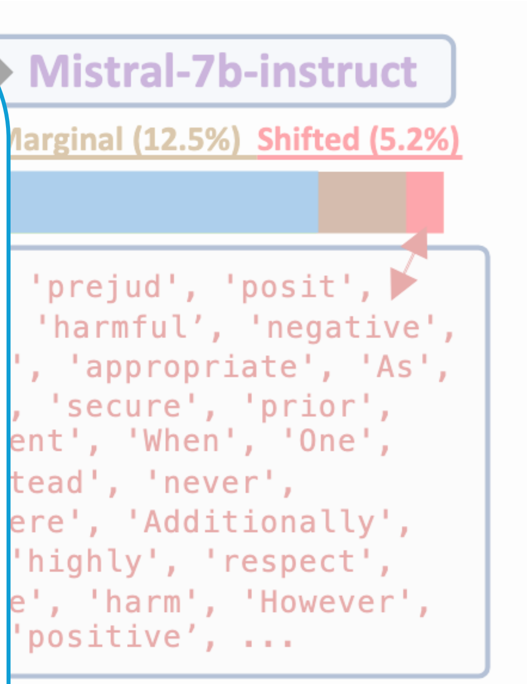


Lin, Bill Yuchen, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. "The unlocking spell on base llms: Rethinking alignment via in-context learning." *arXiv preprint arXiv:2312.01552*

NeuroSymbolic AI for Instructibility

Why Instructibility Needs More Than Instruction Tuning?





- **Token-Level Impact is Minimal**
  - Instruction tuning modifies behavior on only a small fraction of tokens
  - Base and instruction-tuned models generate identical top token choices in most positions
- **Surface-Level Modifications**
  - Instruction tuning represent a tiny fraction of model's overall output
  - Focus is on discourse markers and transitional words
- **Base Model Capability**
  - Given appropriate prompting, base models can match instruction-tuned performance
  - Indicates instruction tuning surfaces existing capabilities rather than teaching new ones

Lin, Bill Yuchen, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. "The unlocking spell on base llms: Rethinking alignment via in-context learning." *arXiv preprint arXiv:2312.01552*

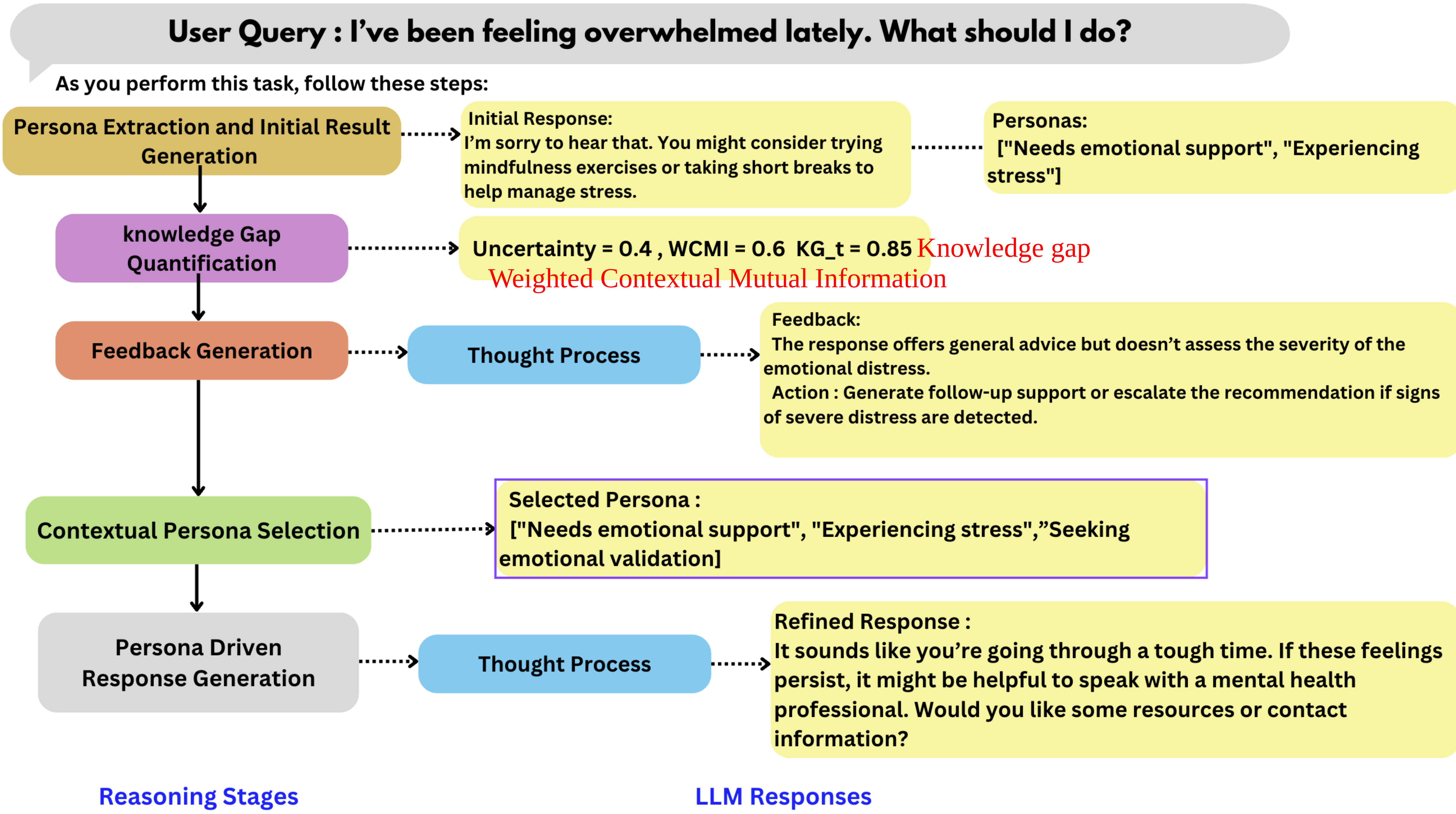
Most Importantly …

🚫 Why This Is Not Enough

•Instruction tuning doesn't fill the knowledge gap; it simply redirects existing behavior.
•Complex domains need deeper reasoning capabilities that go beyond token-level adjustments.

How Instructibility Relates to Knowledge Gaps

- A genuinely instructible LLM must dynamically recognize gaps in its knowledge.
- Requires active reasoning mechanisms to retrieve and apply external knowledge.
- The model must collaborate with external systems (symbolic reasoners, knowledge bases) to fill these gaps effectively.


User Query : I've been feeling overwhelmed lately. What should I do?
As you perform this task, follow these steps:
Persona Extraction and Initial Result Generation
Initial Response:
I'm sorry to hear that. You might consider trying mindfulness exercises or taking short breaks to help manage stress.
Personas:
["Needs emotional support", "Experiencing stress"]
knowledge Gap Quantification
Uncertainty = 0.4 , WCMI = 0.6 KG_t = 0.85
Knowledge gap
Weighted Contextual Mutual Information
Feedback Generation
Thought Process
Feedback:
The response offers general advice but doesn't assess the severity of the emotional distress.
Action : Generate follow-up support or escalate the recommendation if signs of severe distress are detected.
Contextual Persona Selection
Selected Persona :
["Needs emotional support", "Experiencing stress",”Seeking emotional validation]
Persona Driven Response Generation
Thought Process
Refined Response :
It sounds like you're going through a tough time. If these feelings persist, it might be helpful to speak with a mental health professional. Would you like some resources or contact information?
Reasoning Stages
LLM Responses

Automated Metrics

Method	Standard		
	BLEU	ROUGE-L	BERT-F1
Llama 3.1 (0S)	0.001	0.112	0.866
Llama 3.1 (CoT)	0.105	0.139	0.845
Llama 3.1 (SR)	0.001	0.110	0.859
Llama 3.1 (RoT)	0.001	0.103	0.850
Llama 3.1 (Knowledge gap)	0.002	0.128	0.868

Gou, Tian, Boyao Zhang, Zhenglie Sun, Jing Wang, Fang Liu, Yangang Wang, and Jue Wang. "Rationality of thought improves reasoning in large language models." In *International Conference on Knowledge Science, Engineering and Management*, pp. 343-358. Singapore: Springer Nature Singapore, 2024.

Automated Metrics

Method	Standard		
	BLEU	ROUGE-L	BERT-F1

Human & AI Preference Metrics (ESConv)

Method	GPT-pref	Nubia	Human-pref
Llama 3.1 (0S)	3.38%	0.110	17.26%
Llama 3.1 (CoT)	8.47%	0.139	15.77%
Llama 3.1 (SR)	13.56%	0.150	17.86%
Llama 3.1 (RoT)	5.08%	0.128	7.44%
Llama 3.1 (Knowledge gap)	69.49%	0.160	41.66%

LLM Generator ⟷ LLM Evaluator

Conversation Context:

- **Chat History:**
  User: "I've been feeling really overwhelmed with work lately. It's like no matter how hard I try, I can't keep up, and I'm constantly stressed."
  AI: "I'm sorry to hear you're feeling this way. Work stress can be really tough to manage. Have you had a chance to take a break or talk to someone about it?"
  User: "Not really. I just don't feel like I have the time, and I'm not sure it would even help."

User Input:
"What do you think I should do?"

Response Options:

1. **Option 1:**
   "It sounds like you're carrying a lot right now, and that can be so draining. Sometimes, even small breaks can make a difference—maybe just a walk outside or taking five minutes to breathe. Have you tried anything like that before?"
2. **Option 2:**
   "You should try taking a break. Pushing yourself too hard won't help in the long run. Everyone gets overwhelmed sometimes; it's normal. You'll get through it."
3. **Option 3:**
   "I can understand how overwhelming it must feel to juggle so much. Maybe focusing on one small thing you can control right now could help ease the load. Is there a task you could delegate or delay?"
4. **Option 4:**
   "I'm sorry you're feeling this way. Work stress can be really hard to manage. It's important to remember that it's okay to ask for help or take time for yourself."
5. **Option 5:**
   "Stress is part of life, but it's important not to let it control you. Maybe you just need to push through this tough time. You're stronger than you think."

Selecting the Best Response:

When choosing the best response, consider these questions:

1. **Relevance to User Input:**
   - Does the response address the user's feelings of being overwhelmed and their question about what they should do?
   - **Example:** Options 1 and 3 provide actionable suggestions while also acknowledging the user's emotional state.
2. **Conversational Engagement:**
   - Does the response encourage the user to reflect further or engage in a meaningful discussion?
   - **Example:** Option 1 asks a follow-up question about whether the user has tried taking small breaks, fostering engagement.
3. **Contextual Appropriateness:**
   - Does the response fit the emotional tone of the conversation?
   - **Example:** Options 1, 3, and 4 are empathetic and supportive, matching the user's need for emotional support.
4. **Natural Dialogue Flow:**
   - Does the response feel like part of a compassionate and thoughtful conversation?
   - **Example:** Option 1 flows naturally by validating the user's feelings and offering simple, achievable advice.
5. **Persona Alignment:**
   - Does the response align with the persona of an emotionally supportive conversational partner?
   - **Example:** Option 1 aligns well by being empathetic, practical, and warm.
6. **Potential to Continue Interaction:**
   - Does the response invite the user to explore their feelings or thoughts further?
   - **Example:** Options 1 and 3 open the door for more discussion about small steps the user might take.

Automated Metrics

Method	With LLM Evaluator		
	BLEU	ROUGE-L	BERT-F1
Llama 3.1 (0S)	0.008	0.118	0.872
Llama 3.1 (CoT)	0.112	0.145	0.858
Llama 3.1 (SR)	0.015	0.125	0.868
Llama 3.1 (RoT)	0.012	0.128	0.875
Llama 3.1 (Knowledge gap)	0.018	0.142	0.885

Human & AI Preference Metrics (ESConv)

With LLM Evaluator		
GPT-pref	Nubia	Human-pref
5.45%	0.125	19.82%
12.65%	0.148	18.92%
18.92%	0.165	22.45%
82.35%	0.185	48.82%
85.72%	0.192	52.34%

Summary

Definition Integration: The WellDunn framework formalizes the incorporation of clinical definitions into mental health assessment systems, enabling a more accurate understanding of psychological conditions

Rule of Thumb Extraction and Contextualization: SAGE extracts clinical heuristics from mental health knowledge bases as ***rules of thumb***, turning LLM agents into more empathetic and grounded agents.

Semantic Encoding and Decoding Optimization: The SEDO framework preserves nuanced psychological semantics when integrating expert knowledge into mental health assessment systems.

Process Knowledge-infused Learning demonstrates how therapeutic processes and intervention sequences can be incorporated into AI systems to provide ethically sound mental health support.

Knowledge Gaps Assessment enables LLMs to dynamically measure intrinsic contextual uncertainty during conversations, strategically resolving persona knowledge gaps through targeted questions rather than producing hallucinated responses when information is incomplete.

Handoff

Vector Symbolic Architectures

Booz | Allen | Hamilton

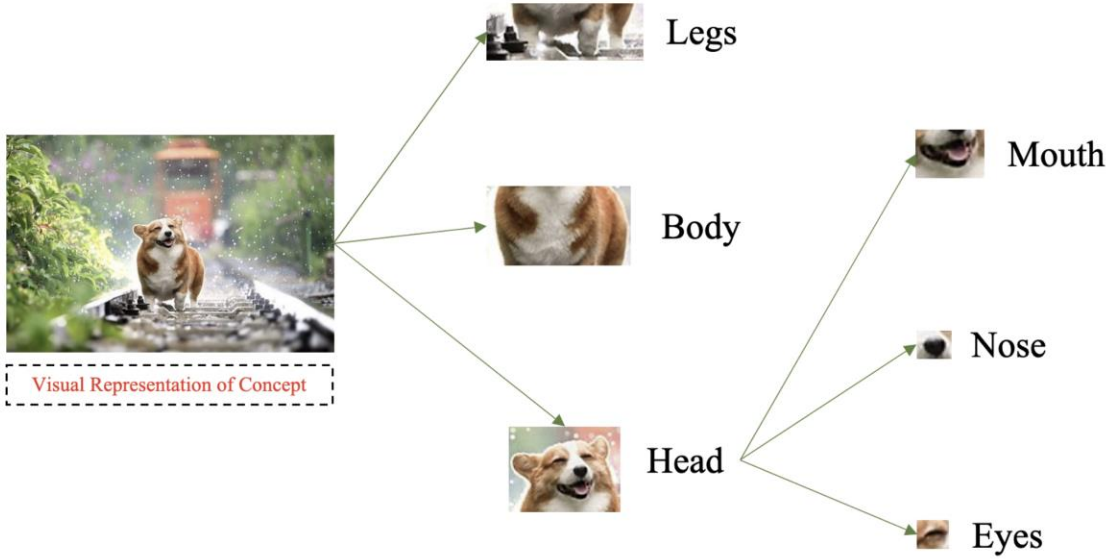
Vector Symbolic architectures in deep learning

Edward Raff

AAAI TUTORIAL NEUROSYMBOLIC AI FOR EGI| 24 FEB 2025

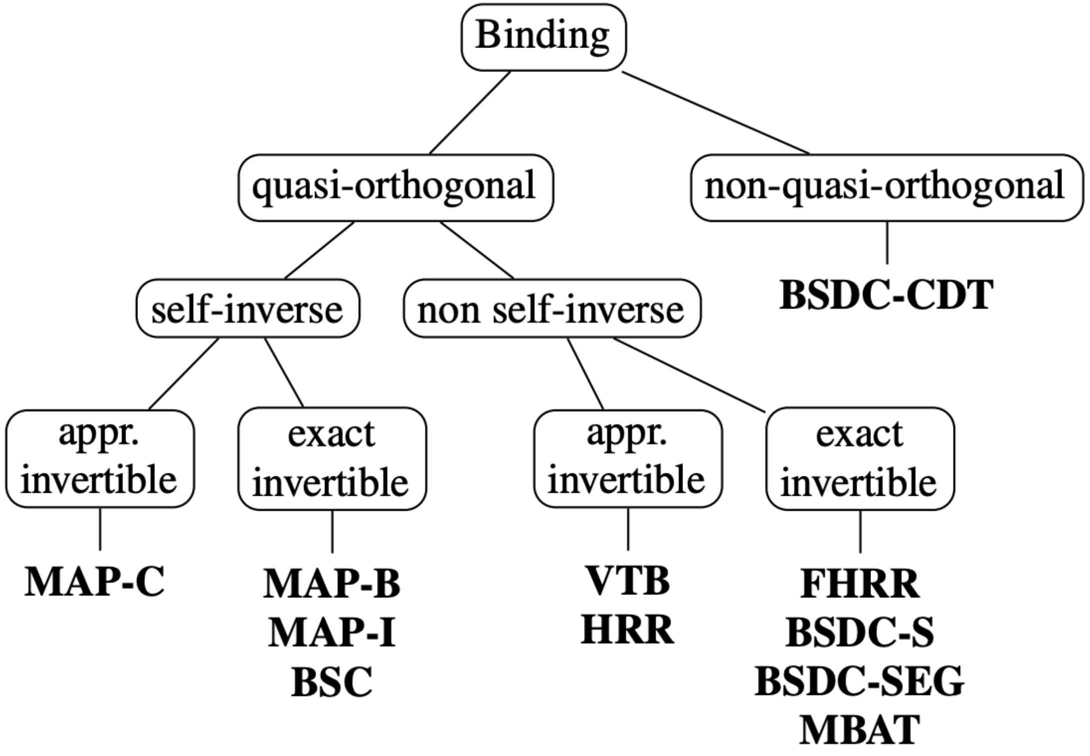
Vector Symbolic Architectures

- Based on "Good Old Fashioned AI" from the 1990s, it is a noisy way to represent symbolic logic.
  - Every "concept" is represented by a vector in *D* dimensional space.
  - Vectors are random, you follow specific rules to create them
  - Because everything is a vector, operations are theoretically differentiable!
- You can manipulate VSA vectors using pre-defined operations.
  - Adding/bundling: $\boldsymbol{x} + \boldsymbol{y}$ you can think of this as an "or" operation.
  - Binding: $\boldsymbol{x} \otimes \boldsymbol{y}$ you can think of this as an "and" operation
  - Permutation: $\varphi(\boldsymbol{x})$ you can think of this as repetition. i
  - Similarity: $\boldsymbol{sim}(\boldsymbol{x}, \boldsymbol{y})$ calculates the dissimilarity between two VSA vectors



VSA operations

- Suppose you have vectors for "red", "cat", "blue" and "dog"
- Property 1: VSAs are quasi-orthogonal and no expected similarity, $\mathrm{sim}(\boldsymbol{x}, \boldsymbol{y}) \approx \boldsymbol{0}$
- Property 2: Binding and permuting results in vectors that dissimilar $\mathrm{sim}(\boldsymbol{x} \otimes \boldsymbol{y}, \boldsymbol{x}) \approx \boldsymbol{0}$ and $\mathrm{sim}(\boldsymbol{\varphi}(\boldsymbol{x}), \boldsymbol{x}) \approx \boldsymbol{0}$
- Property 3: You can invert vectors to unbind, and retrieve similar vectors.
    - $(\boldsymbol{x} \otimes \boldsymbol{y}) \otimes \boldsymbol{y}^{\dagger} \approx \boldsymbol{x}$
    - $(\boldsymbol{x}^{\dagger})^{\top}(\boldsymbol{x} \otimes \boldsymbol{y}) \approx \boldsymbol{1}$
    - $(\boldsymbol{z}^{\dagger})^{\top}(\boldsymbol{x} \otimes \boldsymbol{y}) \approx \boldsymbol{0}$
- Using these properties, we can form complex structures.
    - Red $\otimes$ cat + blue $\otimes$ dog = S, what was red? $(\text{Red}^{\dagger})^{\top}$ S $\approx$ cat
    - Lists: $\mathrm{L} = \varphi(\boldsymbol{p}) \otimes \boldsymbol{x} + \varphi^{\boldsymbol{2}}(\boldsymbol{p}) \otimes \boldsymbol{y} + \cdots$, what was 2nd in the list? $\varphi^{\boldsymbol{2}}(\boldsymbol{p})^{\dagger} \otimes \boldsymbol{L} \approx \boldsymbol{y}$



Holographic Reduced Representations: a primer

- Binding $\boldsymbol{a} \otimes \boldsymbol{b}$ and unbinding with $\boldsymbol{b}^{\dagger} \otimes (\boldsymbol{a} \otimes \boldsymbol{b}) \approx \boldsymbol{a}$ can be done quickly using FFTs. Also called circular convolution
- *But*, the results tend to be numerically unstable. So much so that a pseudo inverse is used almost exclusively in other work.
  - The error from an incorrect inverse is smaller than numerical errors!
- You can force inverse and pseud- inverse to be equal with a projection such that $\pi(\mathbf{a})^{\dagger} = \pi(\mathbf{a})^{*}$ but this is slower.

Binding: $\boldsymbol{a} \otimes \boldsymbol{b} = \mathcal{F}^{-1}(\mathcal{F}(\boldsymbol{a}) \odot \mathcal{F}(\boldsymbol{b}))$

Inverse: $\mathbf{a}^{\dagger} = \mathcal{F}^{-1}\left(\dots, \frac{1}{\mathcal{F}(\boldsymbol{a})_j}, \dots\right)$

Pseudo Inverse: $\boldsymbol{a}^{*} = [a_1, a_d, a_{d-1}, \dots, a_2]$

Initialization: $\boldsymbol{a} \sim \mathcal{N}(0, I_d \cdot d^{-1})$

<u>*Projection*</u>: $\pi(\boldsymbol{a}) = \mathcal{F}^{-1}\left(\dots, \frac{\mathcal{F}(\boldsymbol{a})_j}{|\mathcal{F}(\boldsymbol{a})_j|}, \dots\right)$

Consider a d = 3 dimensional space, where we wish to compute c † ⊗ (c ⊗ x), we will get the result that:

$$\boldsymbol{c}^{\dagger} \otimes (\boldsymbol{c} \otimes \boldsymbol{x}) = \begin{bmatrix} x_0 \boxed{(c_0^2 + c_1^2 + c_2^2)} \boxed{+ x_1 c_0 c_2 + x_2 c_0 c_1 + x_1 c_0 c_1 + x_2 c_1 c_2 + x_1 c_1 c_2 + x_2 c_0 c_2} \\ x_1 \boxed{(c_0^2 + c_1^2 + c_2^2)} \boxed{+ x_0 c_0 c_1 + x_2 c_0 c_2 + x_0 c_0 c_2 + x_2 c_1 c_2 + x_0 c_1 c_2 + x_2 c_0 c_1} \\ x_2 \boxed{(c_0^2 + c_1^2 + c_2^2)} \boxed{+ x_0 c_0 c_1 + x_1 c_1 c_2 + x_0 c_1 c_2 + x_1 c_0 c_2 + x_0 c_0 c_2 + x_1 c_0 c_1} \end{bmatrix}$$

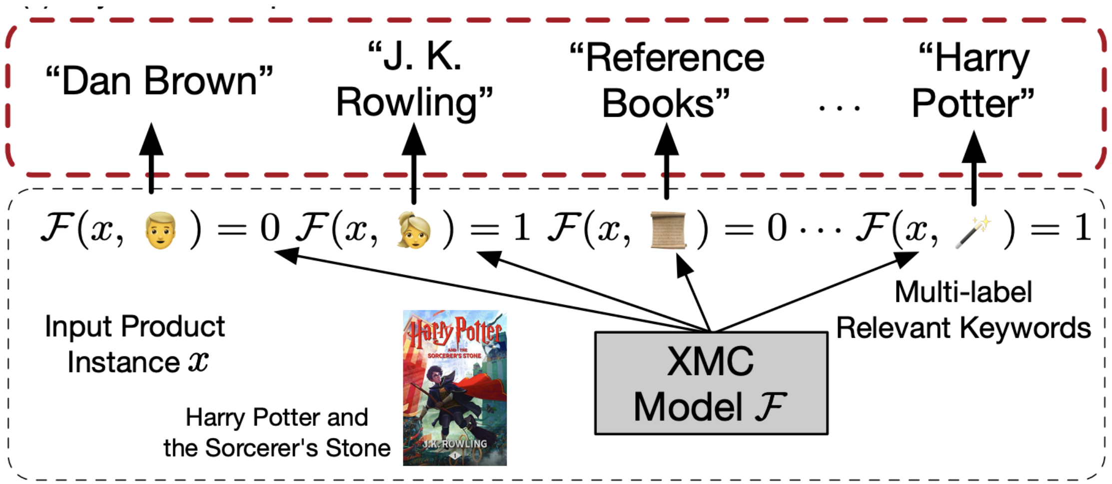
Create $\xi = (c_0^2 + c_1^2 + \cdots + c_d^2) - 1$

$$\boldsymbol{c}^{\dagger} \otimes (\boldsymbol{c} \otimes \boldsymbol{x}) = \begin{bmatrix} x_0(1+\xi) + \eta_0 \\ x_1(1+\xi) + \eta_1 \\ x_2(1+\xi) + \eta_2 \end{bmatrix} = (1+\boldsymbol{\xi})\tilde{\mathbf{x}} + \tilde{\boldsymbol{\eta}}$$

We have $\mathbb{E}[\xi] = 0$ and $\mathbb{E}[\eta_i] = 0$,
so $\mathbb{E}[(1+\boldsymbol{\xi})\tilde{\mathbf{x}} + \tilde{\boldsymbol{\eta}}] = \boldsymbol{x}$

Extreme Multi-label classification (XML)

- We want to develop ways of using VSAs to tackle large output spaces and complexity that we see in malware.
- XML problems have a huge output space, 100,000 to millions of outputs
  - The softmax layer alone is 1-10 GB in size!
  - Most XML papers are trying to come up with ways to break up/factorize this large output space into something more manageable.
- Can we use HRRs to create a compact representation of the output space?
  - Let L be the number of classes/outputs (>100k)
  - We will have an HRR vector $\boldsymbol{c}_{\mathbf{1},\mathbf{2},\cdots,\boldsymbol{L}}$ for each class
  - We will define two special HRR vectors $\boldsymbol{m}$ for “missing” and $\boldsymbol{p}$ for “present”



Product Keyword Recommendation as an XMC task

Eli Chien, Jiong Zhang, Cho-Jui Hsieh, Jyun-Yu Jiang, Wei-Cheng Chang, Olgica Milenkovic, and Hsiang-Fu Yu. 2023. PINA: leveraging side information in extreme multi-label classification via predicted instance neighborhood aggregation. In Proceedings of the 40th International Conference on Machine Learning (ICML'23), Vol. 202. JMLR.org, Article 224, 5616–5630.

A symbolic version of XML

- Represent all classes as $\mathcal{A} = \sum_{i=1}^{L} \boldsymbol{c}_i$ which is $\boldsymbol{\mathcal{O}(L\ d)}$ *once*
- Naïvely we can convert a label to a vector as

$$\mathbf{s} = \overbrace{\sum_{\boldsymbol{c}_p \in \mathcal{Y}^p} \mathbf{p} \otimes \boldsymbol{c}_p}^{\text{Labels present}} + \overbrace{\sum_{\boldsymbol{c}_m \in \mathcal{Y}^m} \boldsymbol{m} \otimes \mathbf{c}_m}^{\text{Labels absent}}$$

- HRRs are distributive! Reduce cost to $\boldsymbol{\mathcal{O}(|\mathcal{Y}^p|\ d)}$ as

$$\boldsymbol{s} = \mathbf{p} \otimes \sum_{\boldsymbol{c}_p \in \mathcal{Y}^p} \boldsymbol{c}_p + \boldsymbol{m} \otimes \left(\mathcal{A} - \sum_{\boldsymbol{c}_p \in \mathcal{Y}^p} \boldsymbol{c}_p \right)$$

- Loss function will be in two parts, one for positives, one for negatives, $\ell = J_p + J_n$

$$J_p = \sum_{\boldsymbol{c}_p \in \mathcal{Y}^p} \left(1 - \cos\left(\mathbf{p}^* \otimes \hat{\mathbf{s}}, \boldsymbol{c}_p\right)\right)$$

$$J_n = \cos\left(\mathbf{m}^* \otimes \hat{\mathbf{s}}, \sum_{c_p \in \mathcal{Y}^p} \boldsymbol{c}_p \right)$$

Table 1: Accuracy of our baseline models and their HRR counterparts with the same network architecture otherwise. Cases where the HRR outperforms its baseline counterpart are in **bold**.

	Bibtex		Delicious		Mediamil		Amazon-12K	
Model	FC	HRR-FC	FC	HRR-FC	FC	HRR-FC	CNN	HRR-CNN
P@1	46.4	**60.3**	65.0	**66.5**	84.8	83.9	89.1	84.5
PSP@1	32.5	**45.6**	64.2	**30.0**	64.2	63.7	49.2	44.2
	EURLex-4K						Amazon-13K	
Model	FC	HRR-FC	CNN	HRR-CNN	LSTM	HRR-LSTM	FC	HRR-FC
P@1	73.4	**77.2**[a]	47.1	**50.0**	63.0	**70.4**	93.0[a]	**93.3**[a,b]
PSP@1	32.0	30.7	18.0	17.5	26.4	**26.8**	52.6	49.6
	Wiki10-31K						Amazon-13K	
Model	FC	HRR-FC	CNN	HRR-CNN	LSTM	HRR-LSTM	LSTM	HRR-LSTM
P@1	80.4[a]	**81.1**[a,b]	60.0	**74.3**	83.5	**85.0**	90.0	**93.4**
PSP@1	9.46	9.19	10.4	9.88	10.6	10.5	48.7	**48.8**
	Delicious-200K		Amazon-670K					
Model	FC	HRR-FC	FC	HRR-FC	CNN	HRR-CNN		
P@1	21.8	**44.9**	34.6[a]	19.9	14.1	6.11		
PSP@1	10.5	6.84	5.22	**8.45**	9.39	1.51		

Smaller models

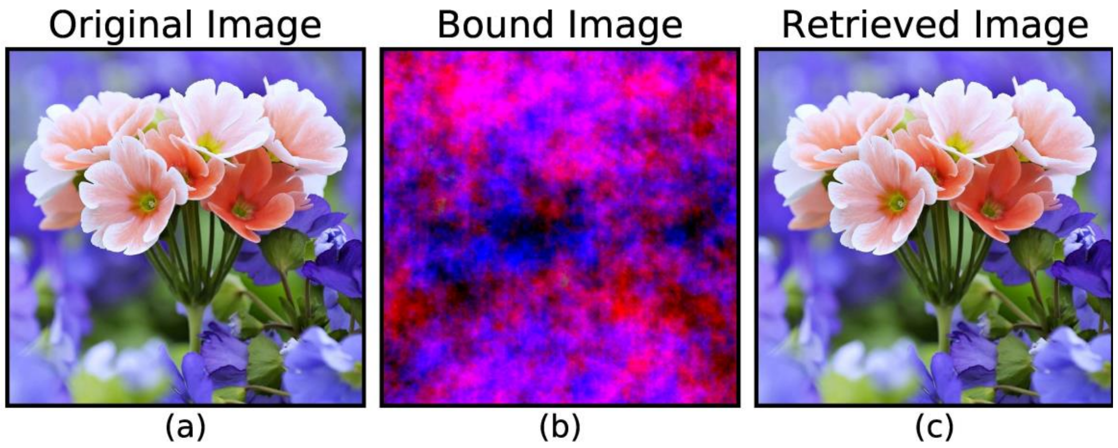
- Accuracy is shockingly sometimes improved, despite massively reducing the network output space.
  - HRR is at a disadvantage, similar classes have disimilar vectors $\boldsymbol{c}_i$
- HRR results in significant smaller models as problem size increases.
  - As little as 1% of the weights compared to the original softmax layers
  - Softmax layer was up to 42% of the original network's total weights!
- HRR had a second disadvantage, less parameters in the network!

Table 2: For each dataset the percentage reduction in parameters of the output layer, and the resulting change for the entire network, by replacing the output layer from a fully-connected softmax with our HRR approach.

Dataset	**Dim d'**	**% Compression**	
		Output	**Network**
Delicious	400	59.30	29.22
EURLex-4K	400	89.98	37.80
Wiki10-31K	3000	90.25	29.49
Amazon-13K	3000	77.49	4.74
Delicious-200K	3000	98.53	41.88
Amazon-670K	3000	99.55	42.09

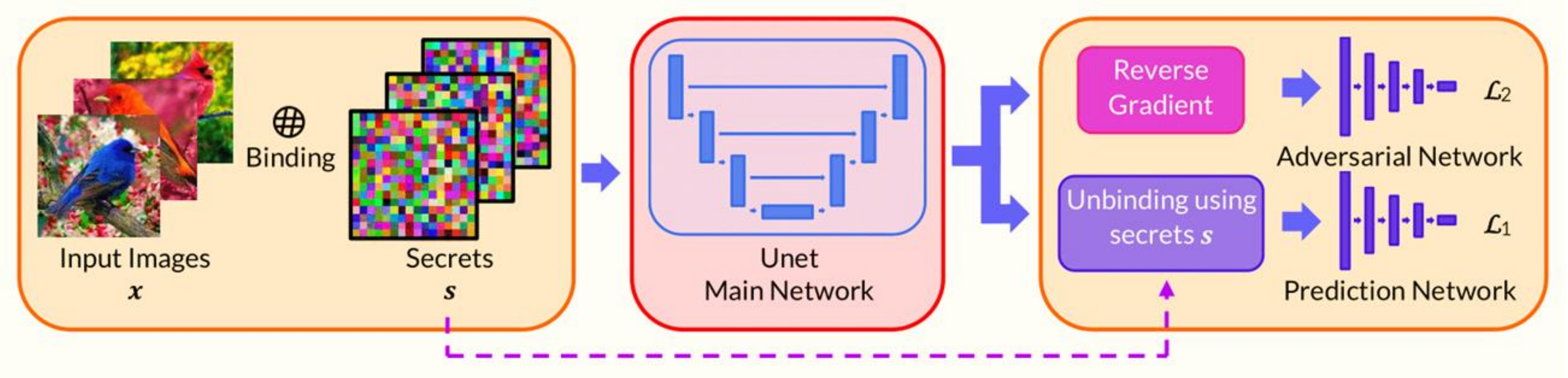
Deploying Convolutional Networks on Untrusted Platforms Using 2D Holographic Reduced Representations

- Say you want to deploy your model on “AWS” to reduce local compute costs.
    - But you don’t fully trust “AWS” to not peak at your data/model, and trying to infer what you are doing.
- Homorphic Encryption is cool, but not actually useful. Far too slow!
- We can exploit the fact that binding inputs are dissimilar between pairs.
- HRR is defined as the result of an FFT, so we can generalize to images via a 2D-FFT



Connectionist Symbolic Pseudo Secrets (CSPS)

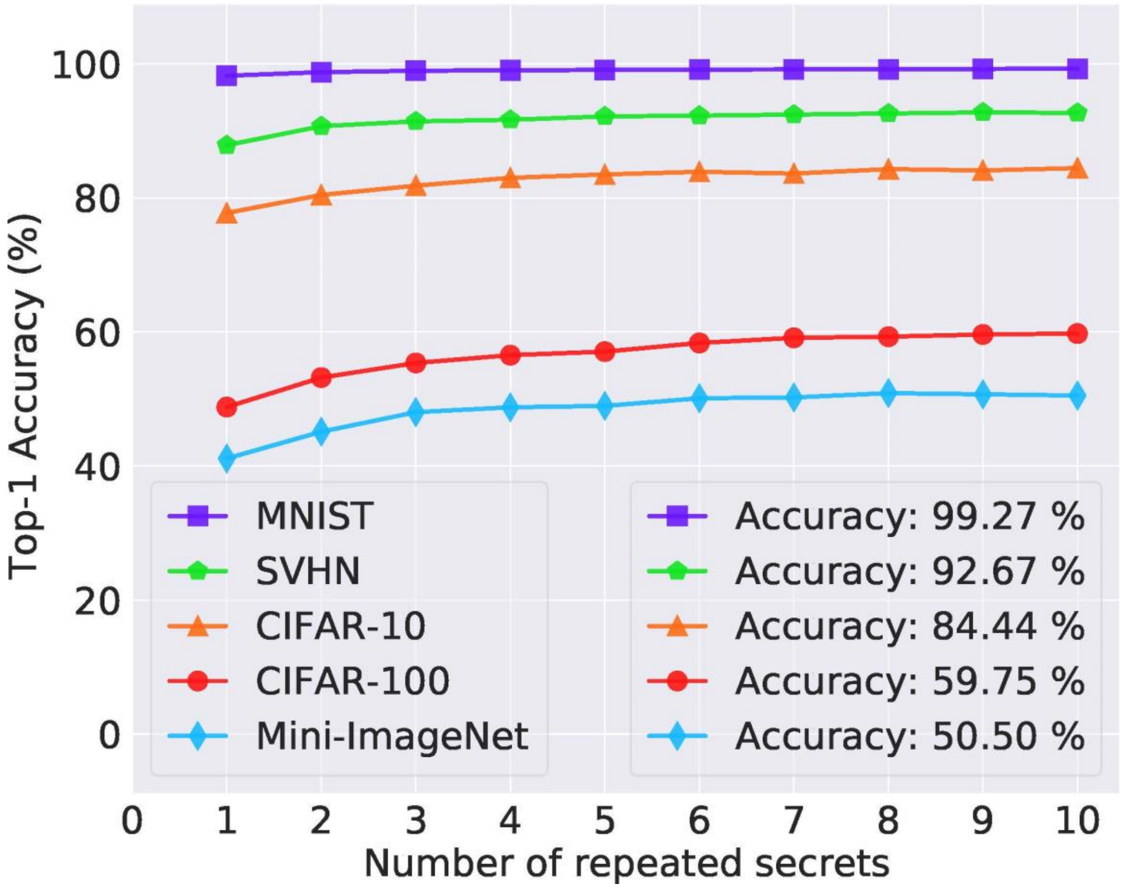
- We are going to mimic a one-time-pad style encryption scheme.
- We will generate a VSA vector $\boldsymbol{s}$ to be our secret known on the local device.
- The "backbone" of the network will be a U-Net run on the untrusted "AWS"
- Two losses during training.
  - $\mathcal{L}_1$ gets access to $\mathbf{z} \otimes \boldsymbol{s}^{\dagger}$ to make the prediction
  - $\mathcal{L}_2$ gets has no access to $\boldsymbol{s}$ to make the prediction
  - Inventive that you need $\boldsymbol{s}$ to make an accurate prediction.



CSPS classification accuracy

- Training is done all locally on one device, it is prediction wea re concerned with.
- Accuracy is reasonable, but some degradation as images get larger.

Dataset	Our CSPS	HE Est.
MNIST	4.56 Seconds	2 Hours 46 Minutes
SVHN	12.44 Seconds	55 Hours 32 Minutes
CIFAR-10	7.58 Seconds	21 Hours 20 Minutes
CIFAR-100	9.07 Seconds	43 Hours 53 Minutes
Mini-ImageNet	28.37 Seconds	Timeout



Do you get any data saving?

- CSPS can offload at least 65% of computation to a remote machine effectively
  - Some work needs to be done locally for unbind/binding, and we need a small classification head to run locally (too hard for a “linear probe”)

Dataset	Remote %	Local %
MNIST	74.24	25.76
SVHN	65.06	34.94
CIFAR-10	66.08	33.92
CIFAR-100	66.78	33.22
Mini-ImageNet	74.42	25.58

Are we giving away any secrets?

- Utility is predicated on obscurity!
    - Can the adversary ease drop on the input and tell what is happening? The output?
- Use clustering algorithms, with the true value of the number of classes, to see if they can find the signal. Adjusted Rand index should be =0 for purely random clustering. Far below 100%!

		MNIST	SVHN	CIFAR-10	CIFAR-100	MiniImgNet
Network's Output	K-Means	1.28	0.06	0.21	0.03	0.08
	Spectral	0.01	0.01	0.00	0.00	0.02
	GMM	1.28	0.06	0.17	0.04	0.09
	Birch	1.51	0.03	0.13	0.05	0.07
	HDBSCAN	0.00	0.00	0.00	0.00	0.00
Network's Input	K-Means	-0.02	-0.01	0.18	0.54	0.42
	GMM	0.01	0.00	0.09	0.61	0.44
	Birch	0.20	0.00	0.14	0.45	0.35
	HDBSCAN	0.00	-0.24	1.23	0.01	0.02

What if the adversary was even smarter

- Using CSPS data with true labels to train a model directly requires an impossibly strong adversary.
- Barely beats random-guessing performance, even with an impossible advantage, is a good sign for heuristic security.
  - MNIST, SVHN, and CIFAR-10 accuracies are at best 2.1× better than random-guessing
  - For CIFAR-100, and Mini-ImageNet it is 2.6 and 4.7× of random guessing

Dataset	Top-1 Accuracy (%)	Top-5 Accuracy (%)	Random Guess
MNIST	19.72	–	10%
SVHN	21.13	–	10%
CIFAR-10	12.91	–	10%
CIFAR-100	2.66	10.33	1%
Mini-ImageNet	4.68	15.01	1%

WHAT ABOUT USING ADVERSARIAL ML?

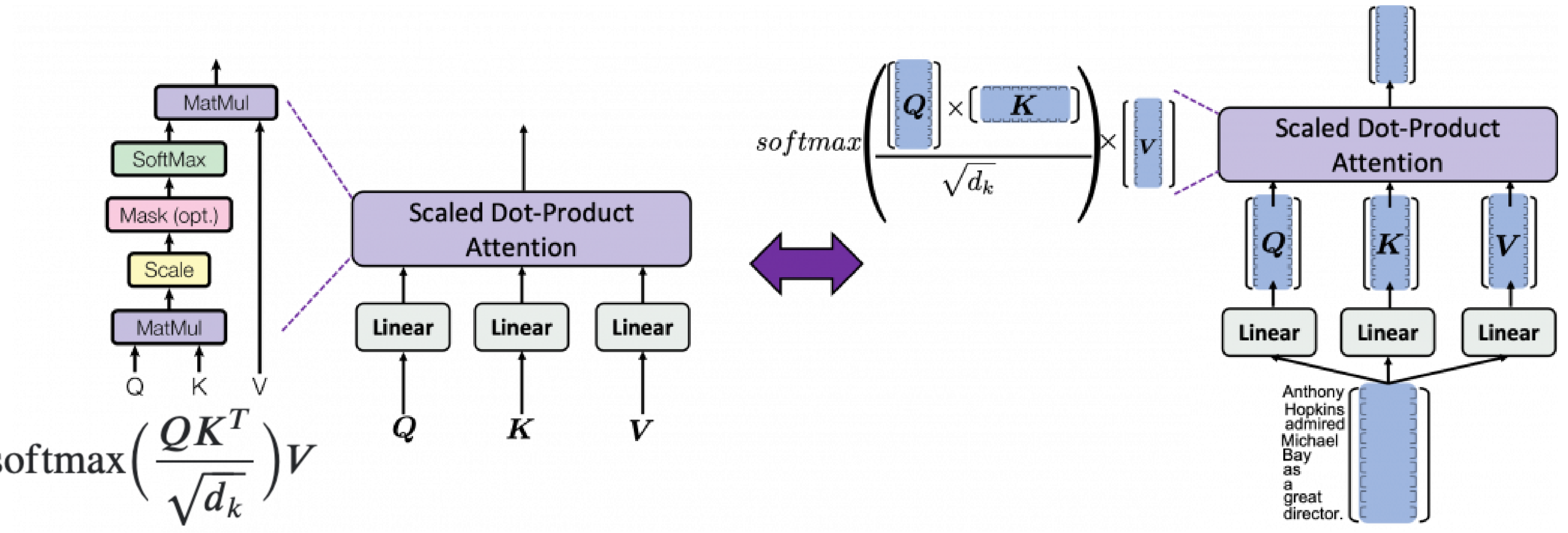
Adversary has access to some of the original images, can it learn the secret using projected gradient descent (PGD)?

WHAT ABOUT TRAINING AN AUTO-ENCODER?

Adversary has access to some of the original images, can it learn by training an Autoencoder?

Self Attention

- Construct a larger sentence $S = \sum_{i=1}^{n} \boldsymbol{x}_i \otimes \boldsymbol{y}_i$.
- If $\mathbf{a} \otimes \boldsymbol{b} \in S$, then $\mathbf{a}^T\left(S \otimes \boldsymbol{b}^{\dagger}\right) \approx 1$. If $\mathbf{a} \otimes \boldsymbol{b} \notin S$, then $\mathbf{a}^T\left(S \otimes \boldsymbol{b}^{\dagger}\right) \approx 0$
- We can test this for varying number of terms bound together. We can see our projection is far more stable, allowing accurate use for far larger number of over 1000 pairs of vectors symbolically contained in a 256-dimensional space!
- This also solves the backpropagation issues, allowing us to learn via gradient descent when using HRRs.



$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Self Attention with HRRs

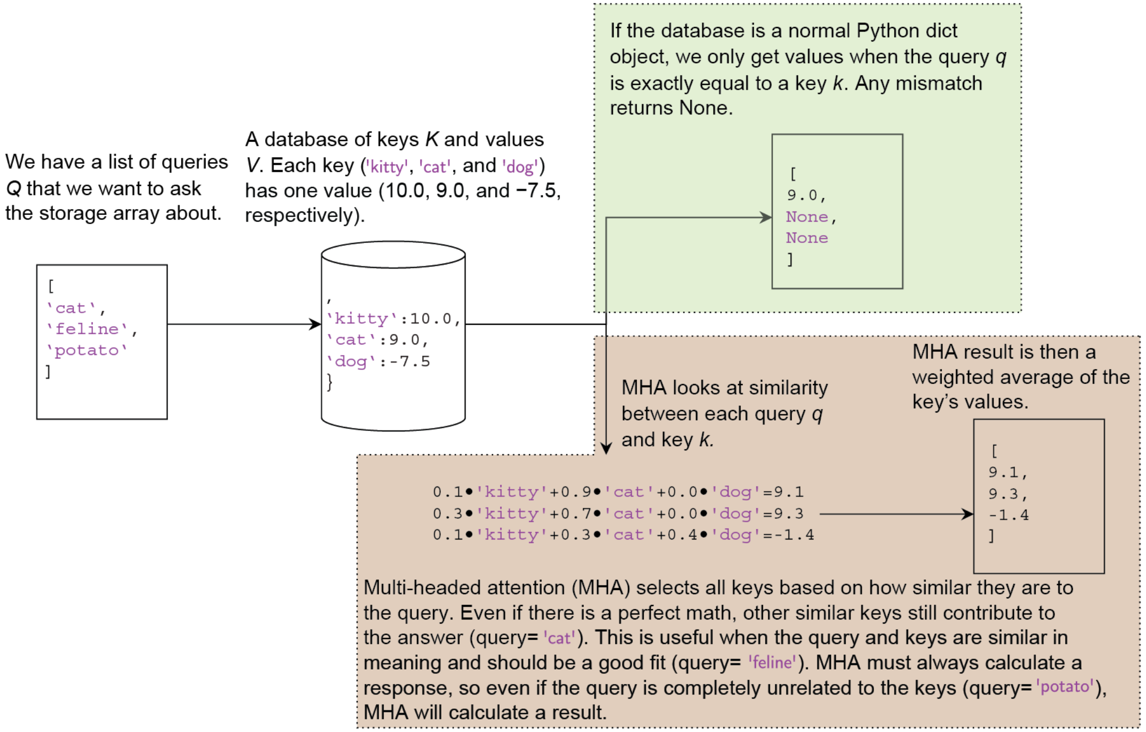
We can think of Self-Attention as a fuzzy dictionary. We are finding the match between query and key and returning the corresponding value – made fuzzy by averaging based on similarity.

Construct a "dictionary" $\beta = \sum_{i=1}^{n} \boldsymbol{k}_i \otimes \boldsymbol{v}_i$.

If $\mathbf{a} \otimes \boldsymbol{b} \in S$, *then* $\mathbf{a}^{\mathsf{T}}\left(\beta \otimes \boldsymbol{b}^{\dagger}\right) \approx 1$.

If $\mathbf{a} \otimes \boldsymbol{b} \notin \beta$, *then* $\mathbf{a}^{\mathsf{T}}\left(\beta \otimes \boldsymbol{b}^{\dagger}\right) \approx 0$

So we can perform querying against this "dictionary", using HRRs as an inductive biased toward key/value lookups!



Self Attention with HRRs: Implementation

1. *Combine all key-value pairs*

$$\beta = \sum_{\{i=1\}}^{T} \boldsymbol{k}_i \otimes \boldsymbol{v}_i$$

2. *query for multiple items at once*

$$\widehat{\boldsymbol{v}} = \left(\sum_{j=1} \boldsymbol{q}_j\right)^{\dagger} \otimes \boldsymbol{\beta}$$

3. *Compute weights of values with*

$$\alpha_t = cossim(\boldsymbol{v}_t, \widehat{\boldsymbol{v}})$$

4. *Normalize with* $\boldsymbol{w} = softmax(\boldsymbol{\alpha})$

5. *Compute final* $Attention(Q, K, V) = [w_1\boldsymbol{v}_1, w_2\boldsymbol{v}_2, \dots, w_T\boldsymbol{v}_T]$

We call the transformer using our HRR self attention a "Hrrformer"

```
class MultiHeadedAttention(nn.Module):
    features: int
    heads: int = 8

    def setup(self):
        self.binding = Binding()
        self.unbinding = Unbinding()
        self.similarity = CosineSimilarity()

    @nn.compact
    def __call__(self, query, key, value, mask=None):
        dense_q = nn.Dense(features=self.features)
        dense_k = nn.Dense(features=self.features)
        dense_v = nn.Dense(features=self.features)
        dense_o = nn.Dense(features=self.features)

        q = dense_q(query)  # (B, T, H)
        k = dense_k(key)    # (B, T, H)
        v = dense_v(value)  # (B, T, H)

        q = split(q, self.heads)  # (B, h, T, H')
        k = split(k, self.heads)  # (B, h, T, H')
        v = split(v, self.heads)  # (B, h, T, H')

        bind = self.binding(k, v)                          # (B, h, T, H')
        bind = np.sum(bind, axis=-2, keepdims=True)  # (B, h, 1, H')

        vp = self.unbinding(bind, q)                                      # (B, h, T, H')
        score = self.similarity(v, vp, axis=-1, keepdims=True)  # (B, h, T, 1)

        if mask is not None:
            score = score + (1. - mask) * (-1e9)  # (B, h, T, 1)
        weight = nn.softmax(score, axis=-2)       # (B, h, T, 1)
        weighted_value = weight * v               # (B, h, T, H')

        output = merge(weighted_value)  # (B, T, H)
        output = dense_o(output)        # (B, T, H)
        return output                   # (B, T, H)
```

Self Attention with HRRs: Noise?

Our self attention works without Gaussian IID coefficients, how? Consider the H dimensional vectors a, b, c, d, and z. If each element of all these vectors is sampled from N (0, 1/H), then we would expect that $(a \otimes b + c \otimes d)^T a^\dagger \approx 1$.Similarly, the value z is not present, so we expect that $(a \otimes b + c \otimes d)^T z^\dagger \approx 0$. Now let's pretend we have 2D data:

We can query for **a** + **z** get and get

$$\frac{(a_0 + z_0)\left(a_0 b_0 + a_1 b_1 + c_0 d_0 + c_1 d_1\right) - (a_1 + z_1)\left(a_0 b_1 + a_1 b_0 + c_0 d_1 + c_1 d_0\right)}{(a_0 - a_1 + z_0 - z_1)(a_0 + a_1 + z_0 + z_1)}$$

Or we can do **c** + **z** and get:

$$\frac{(c_0 + z_0)\left(a_0 b_0 + a_1 b_1 + c_0 d_0 + c_1 d_1\right) - (c_1 + z_1)\left(a_0 b_1 + a_1 b_0 + c_0 d_1 + c_1 d_0\right)}{(c_0 - c_1 + z_0 - z_1)(c_0 + c_1 + z_0 + z_1)}$$

In either case, the noise terms share many coefficents, and will result in similar magnitude noise. We can interpret this as an additional noise constant ε that we must add to each noise term. Then when we apply the softmax operation, we obtain the benefit that the softmax function is invariant to constant shifts in the input, i.e., $\forall \varepsilon \in \mathbb{R}$, softmax(x + ε) = softmax(x). Thus, our softmax *effectively acts as a clean-up operation over the original values!*

Long Range Arena Results

- We use each datasets. Feature Vectors for all datasets were available, so all have FC results.
- Most methods are worse than a naïve Transformer, but Hrrformer is always better!

Table 1: Accuracy results of Hrrformer on Long Range Arena (LRA) benchmark. Our Multi-layer results use the same layer count (3-6) per task as prior methods. Even using just one layer Hrrformer is highly competitive, and the only method besides Luna to be a Pareto improvement over the original Transformer. Our method is further advantaged in that it requires $10\times$ fewer epochs to reach competitive accuracies.

Model	ListOps (2k)	Text (4k)	Retrieval (4k)	Image (1k)	Path (1k)	Path-X (16k)	Avg	Epochs
Transformer (Vaswani et al. 2017)	36.37	64.27	57.46	42.44	71.40	FAIL	54.39	200
Local Attention (Tay et al. 2020c)	15.82	52.98	53.39	41.46	66.63	FAIL	46.06	200
Linear Transformer (Katharopoulos et al. 2020)	16.13	65.90	53.09	42.34	75.30	FAIL	50.55	200
Reformer (Kitaev, Kaiser, and Levskaya 2020)	37.27	56.10	53.40	38.07	68.50	FAIL	50.67	200
Sparse Transformer (Child et al. 2019)	17.07	63.58	59.59	44.24	71.71	FAIL	51.24	200
Sinkhorn Transformer (Tay et al. 2020b)	33.67	61.20	53.83	41.23	67.45	FAIL	51.29	200
Linformer (Wang et al. 2020)	35.70	53.94	52.27	38.56	76.34	FAIL	51.36	200
Performer (Choromanski et al. 2020)	18.01	65.40	53.82	42.77	77.05	FAIL	51.41	200
Synthesizer (Tay et al. 2020a)	36.99	61.68	54.67	41.61	69.45	FAIL	52.88	200
Longformer (Beltagy, Peters, and Cohan 2020)	35.63	62.85	56.89	42.22	69.71	FAIL	53.46	200
BigBird (Zaheer et al. 2020)	36.05	64.02	59.29	40.83	74.87	FAIL	55.01	200
F-Net (Lee-Thorp et al. 2021)	35.33	65.11	59.61	38.67	*77.78*	FAIL	54.42	200
Nystromformer (Xiong et al. 2021)	37.15	65.52	**79.56**	41.58	70.94	FAIL	58.95	200
Luna-256 (Ma et al. 2021)	37.98	65.78	79.56	47.86	**78.55**	FAIL	**61.95**	200
H-Transformer-1D (Zhu and Soricut 2021)	**49.53**	**78.69**	63.99	46.05	68.78	FAIL	61.41	200
Hrrformer Single-layer	38.86	66.49	75.13	47.90	72.79	FAIL	60.23	**20**
Hrrformer Multi-layer	38.24	65.90	75.83	**48.41**	73.17	FAIL	60.31	**20**

Interpretability

- We can visualize the attention weights that our model uses for each prediction, and see if they correspond with the content of the image.
- In doing so, we see that the attention maps precisely to informative outlines/content of the image.
- Remember: the task is linearized images! So the model is learning 2D structure from 1D representations!

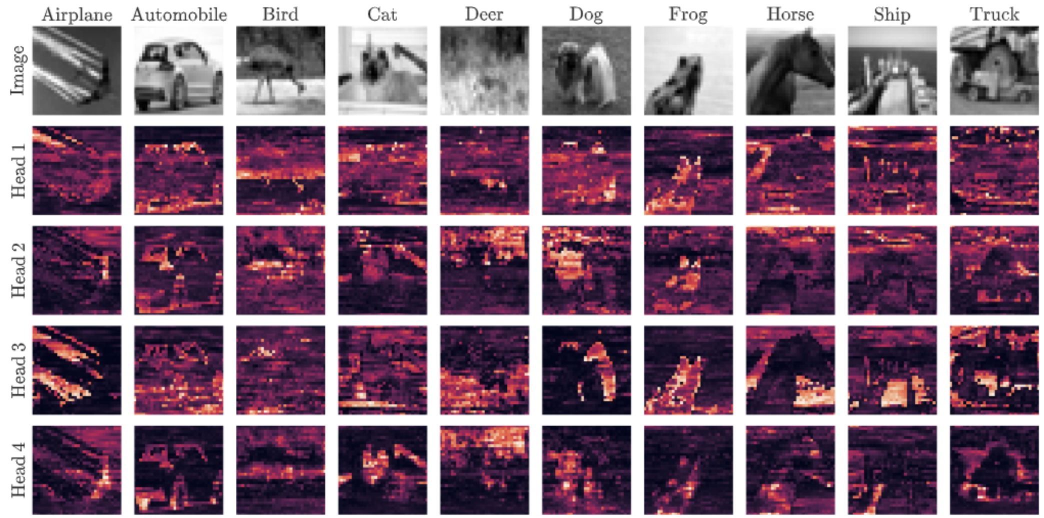


Figure 1: Visualization of weight vector $\mathbf{w} \in \mathbb{R}^{1024 \times 1}$ reshaped to 32×32, the shape of the original image of the CIFAR-10 dataset used in the LRA Image classification task. A single layer Hrrformer is able to learn the 2D structure from the 1D sequence of the image. This is particularly noticeable in the Airplane, dog, Frog, and Horse images. Note context sensitive Head activation can be observed comparing Head 3 for dog vs Frog, where activation occurs for different pixel intensities indicating the model is not naively activating for simple color intensity.

Fast & Low Memory Training

- Transformers are being used in ever larger and more expensive models. Are we fighting that trend? Yes!
- Hrrformer is the fastest by far compared to all prior methods, up to 2 orders of magnitude.
- The Hrrformer uses less memory to train by an order of magnitude or more, depending on what baseline we compare against
- It is nearly the most accurate on the LRA benchmarks.

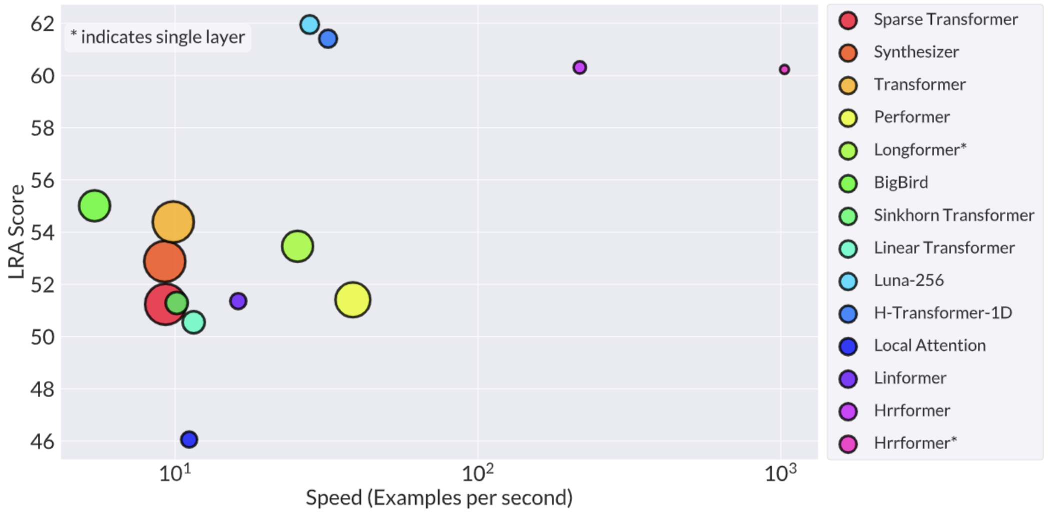


Figure 2: Performance (y-axis), Speed (x-axis, log-scale) of different xformers, and memory footprint on GPU are illustrated by the size of the circles. Hrrformer is in the top-right of the graph, with the smallest circle size, indicating it is the fastest and most memory efficient for training (this does *not* factor in convergence speed).

Fast Predictions

- The FFT function has better numerical behavior as a function of batch size than matrix multiplication.
- The gap between time/sample is smaller for varying batch sizes
- Even our worst case time is better than a transformer's best-case time!

Batch Size	Hrrformer Time (s)	Transformer Time (s)
1	48.21	4429.64
2	33.31	2231.20
3	33.46	1501.59
4	32.32	1125.23
5	30.73	911.32
6	29.76	768.42
7	29.02	661.68
8	28.20	571.63
9	27.78	513.04
10	26.94	463.29
11	26.30	428.68
12	25.80	397.56
13	25.20	366.93
14	24.94	342.73
15	23.87	322.05
16	24.14	305.88
17	24.20	289.57
18	23.71	276.32
19	23.71	265.57
20	23.48	254.24
21	22.98	244.57
22	23.56	236.27
23	22.57	228.23
24	23.21	221.07
25	22.66	216.19
26	22.58	209.43
27	22.14	203.78
28	22.95	199.20
29	22.72	197.80
30	22.24	187.72
31	22.31	183.33
32	21.86	178.31

Malware results

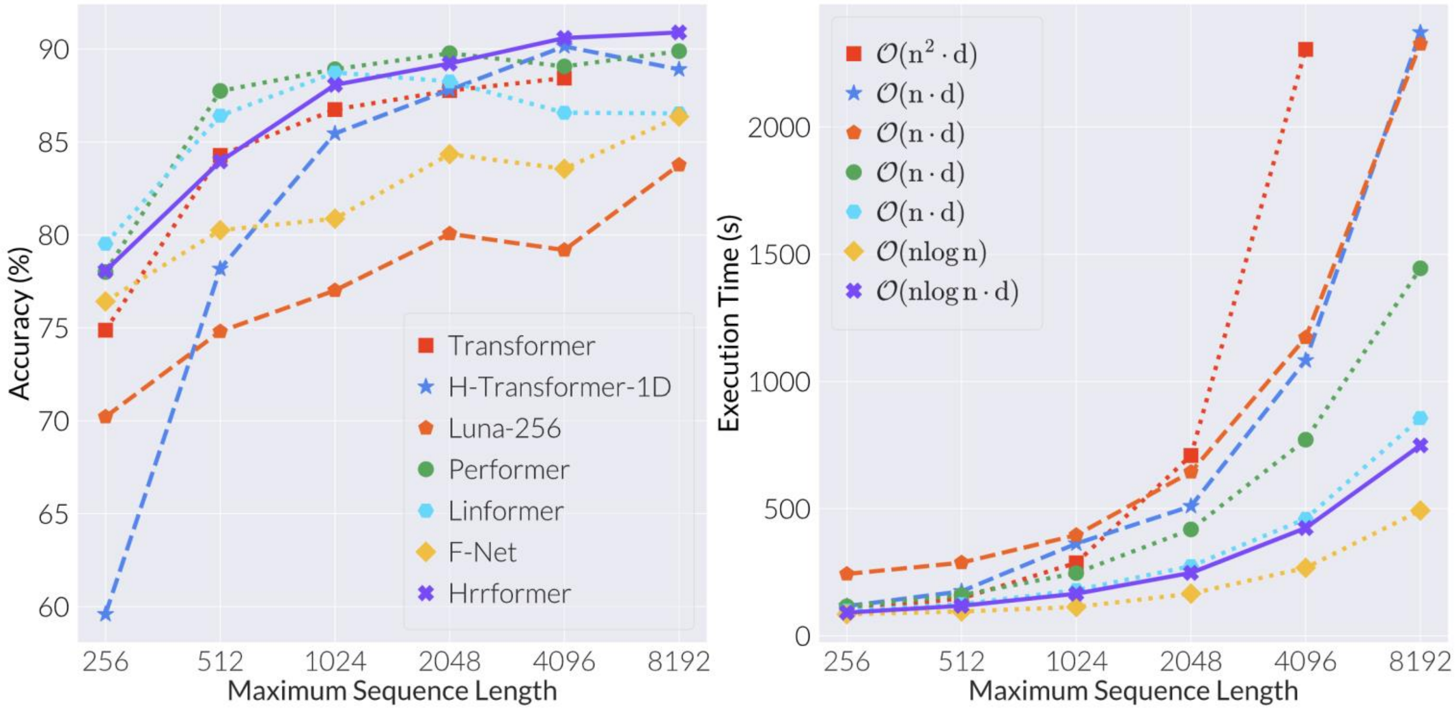


Figure 3: Comparison of Hrrformer with other self-attention models in EMBER malware classification dataset. Hrrformer is presented in a *solid* line and achieves the best accuracy, and second fastest run-time. The two prior best models according the the Long Range Arena H-Transformer-1D and Luna-256 are in the *dashed* lines, and do not perform as well as the LRA would have indicated in speed or accuracy. The rest of the models are in the *dotted* line. This shows the Hrrformer is one of the best options in benchmarks and real-world tasks.

A Walsh Hadamard Derived Linear Vector Symbolic Architecture

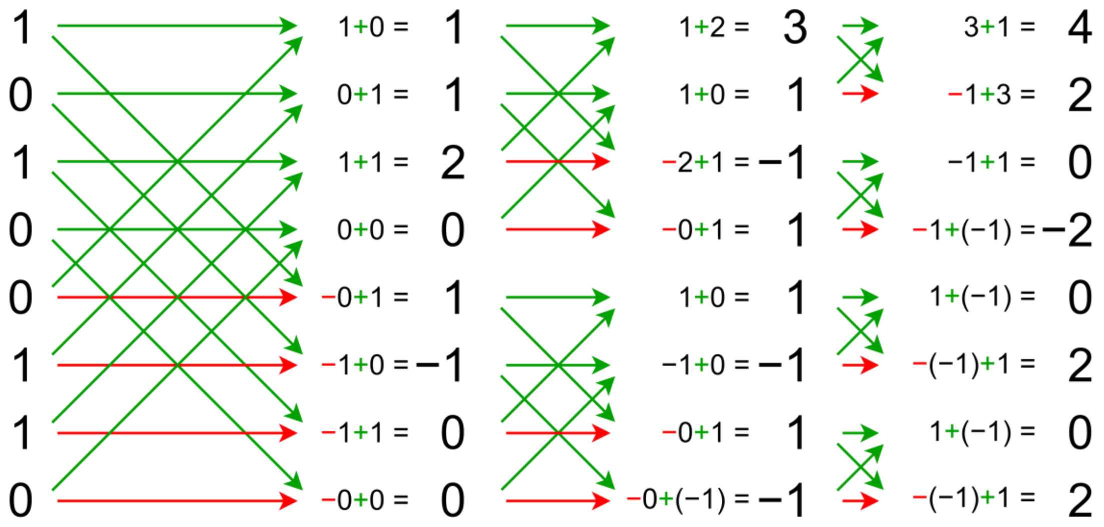
- HRR can be derived by setting up the equation for the identity function, $\mathcal{F}\left(x^{\dagger}\right)_i \mathcal{F}(x)_i = 1$, and solving for what the † operation shoul[d be]
- What ... derive...
  - Avo... aved in g...
- We us... he Hada...

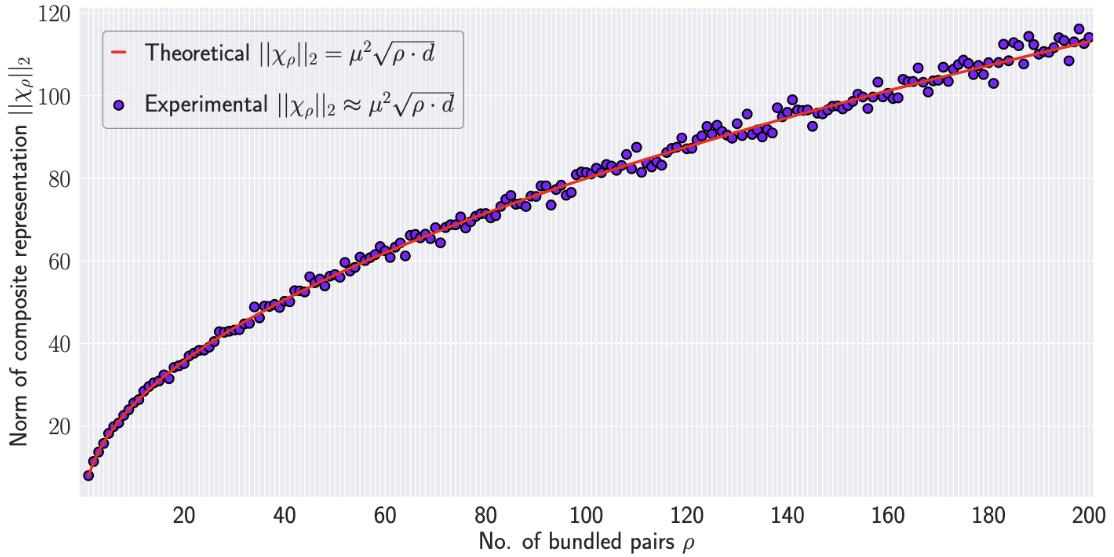
Table 1: The binding and initialization mechanisms for our new HLB with baseline methods. HLB is related to the HRR in being derived via a similar approach, but replacing the Fourier transform $\mathcal{F}(\cdot)$ with the Hadamard transform (which simplifies out). The MAP is most similar to our approach in mechanics, but the difference in derived unbinding steps leads to dramatically different performance. The VTB is the most recently developed VSA in modern use. The matrix $V_{\boldsymbol{y}}$ of VTB is a block-diagonal matrix composed from the values of the $\boldsymbol{y}$ vector, which we refer the reader to [11] for details. The TorchHD library [14] is used for implementations of prior methods.

Method	Bind $\mathcal{B}(x, y)$	Unbind $\mathcal{B}^*(x, y)$	Init x
HRR	$\mathcal{F}^{-1}(\mathcal{F}(\boldsymbol{x}) \odot \mathcal{F}(\boldsymbol{y}))$	$\mathcal{F}^{-1}(\mathcal{F}(\boldsymbol{x}) \oslash \mathcal{F}(\boldsymbol{y}))$	$x_i \sim \mathcal{N}(0, 1/d)$
VTB	$V_y x$	$V_y^\top x$	$\tilde{x}_i \sim \mathcal{N}(0,1) \rightarrow x = \tilde{\boldsymbol{x}} / \lVert \tilde{\boldsymbol{x}} \rVert_2$
MAP-C	$x \odot y$	$x \odot y$	$x_i \sim \mathcal{U}(-1, 1)$
HLB	$x \odot y$	$x \oslash y$	$x_u \sim \{\mathcal{N}(-\mu, 1/d), \mathcal{N}(\mu, 1/d)\}$

Properties of the HLB

- Several useful properties for developing with a VSA
- $\mathbb{E}[\boldsymbol{x}] = 0,$
- $\mathbb{E}[|\boldsymbol{x}|] = \mu$
- $\mathbb{E}\left[\left|\left|\boldsymbol{x}\right|\right|_2\right] = \sqrt{\mu^2 d}$
- If ρ items have been bound together, then$\mathbb{E}[\text{sim}\mathbb{E}[\boldsymbol{x}] \cdot \rho] = 1$ if $\boldsymbol{x} \in \boldsymbol{S}$ and 0 otherwise
  - If you don't know ρ, you can estimate it as $\rho \approx \sqrt{\mu^4 d}$





Good at classical VSA tasks

- When binding with random or repeated VSA vectors, HLB's similarity score remains constant.
- The magnitude of the vector does not change either as more items are bound
- Better to design around something that has a known response

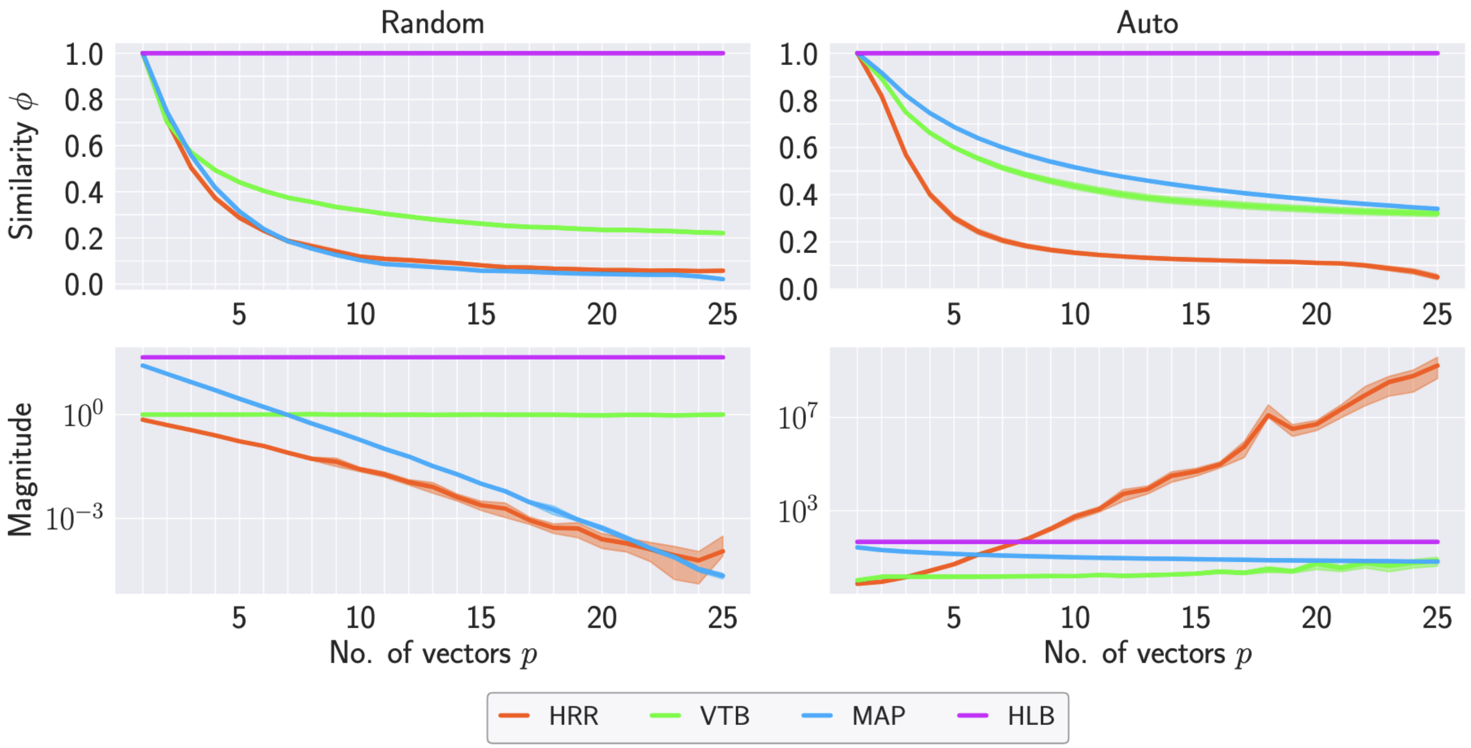


Figure 3: When repeatedly binding different random (left) or a single vector (right), HLB consistently returns the ideal similarity score of 1 for a present item (top row) and has a constant magnitude (bottom row), avoiding exploding/vanishing values.

Better at XML classification

Table 4: XML classification results in dense label representation with HRR, VTB, MAP, and HLB in terms of nDCG and PSnDCG. The proposed HLB has attained the best nDCG and PSnDCG scores on all the datasets setting a new SOTA.

Dataset	Bibtex		Delicious		Mediamill		EURLex-4K	
Metrics	nDCG	PSnDCG	nDCG	PSnDCG	nDCG	PSnDCG	nDCG	PSnDCG
HRR	60.296	45.572	66.454	30.016	83.885	63.684	77.225	30.684
VTB	57.693	45.219	63.325	31.449	87.232	66.948	76.964	31.180
MAP	59.280	46.092	65.376	31.943	87.255	66.886	72.439	26.752
HLB	**61.741**	**48.639**	**67.821**	**32.797**	**88.064**	**67.525**	**77.868**	**31.526**

Dataset	EURLex-4.3K		Wiki10-31K		Amazon-13K		Delicious-200K	
Metrics	nDCG	PSnDCG	nDCG	PSnDCG	nDCG	PSnDCG	nDCG	PSnDCG
HRR	84.497	38.545	81.068	9.185	93.258	49.642	44.933	6.839
VTB	84.663	38.540	78.025	9.645	92.373	49.463	44.092	6.664
MAP	85.472	39.233	80.203	10.027	92.013	48.686	45.373	6.862
HLB	**88.204**	**43.622**	**83.589**	**11.869**	**93.672**	**50.270**	**46.331**	**6.952**

Does better at CSPS

- HLB is more accurate at the CSPS task than prior VSAs
- Also better at hiding its information for CSPS too!
- Because CSPS is purely elementwise operations, no extra work for 2D/n-D generalization

Table 3: Clustering results of the main network inputs (top rows) and outputs (bottom rows) in terms of Adjusted Rand Index (ARI). Because CSPS is trying to hide information, scores near zero are better. Cell color corresponds to the cell absolute value, with blue indicating lower ARI and red indicating higher ARI. All numbers in percentages, and show HLB is better at information hiding.

Clustering Methods	HRR					VTB				
	MNIST	SVHN	CR10	CR100	MIN	MNIST	SVHN	CR10	CR100	MIN
K-Means	−0.02	−0.01	0.18	0.54	0.42	−0.00	−0.01	−0.01	0.02	0.00
GMM	0.01	0.00	0.09	0.61	0.44	4.67	1.37	−0.01	0.02	0.01
Birch	0.20	0.00	0.14	0.45	0.35	0.02	0.03	0.04	0.08	0.03
HDBSCAN	0.00	−0.24	1.23	0.01	0.02	0.00	0.00	0.00	0.00	0.00
K-Means	1.28	0.06	0.21	0.03	0.08	8.52	0.13	1.11	0.05	0.12
GMM	1.28	0.06	0.17	0.04	0.09	8.63	0.14	1.63	0.05	0.00
Birch	1.51	0.03	0.13	0.05	0.07	3.24	0.00	0.64	0.06	0.17
HDBSCAN	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.00	0.00	0.00

Clustering Methods	MAP					HLB				
	MNIST	SVHN	CR10	CR100	MIN	MNIST	SVHN	CR10	CR100	MIN
K-Means	0.17	0.01	0.01	0.00	0.00	0.09	0.00	0.00	0.00	0.00
GMM	3.39	−0.01	0.01	0.00	0.00	2.53	0.00	0.00	0.00	0.00
Birch	0.84	−0.00	0.00	0.01	0.00	0.83	0.00	0.00	0.01	0.00
HDBSCAN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
K-Means	15.91	0.09	0.00	0.03	0.01	13.67	−0.04	0.01	0.02	−0.00
GMM	42.43	0.11	0.00	0.03	0.00	14.96	−0.04	0.01	0.02	0.00
Birch	7.09	−0.07	−0.02	0.01	−0.00	18.44	−0.07	0.00	0.01	0.02
HDBSCAN	0.48	0.00	0.00	0.00	0.00	7.60	0.01	0.00	0.00	0.00

Questions?

We can use VSAs to create neuro-symbolic AI methods

- We can be clever with the loss function to impose symbolic constraints
- We can design layers with symbolic interpretations as a way to express priors
- We can design and simplify our approach to VSAs to achieve better results

Edward Raff
EdwardRaff.com
Raff_Edward@bah.com


Handoff

Grounding Blackbox Language Models with Retrieval Augmented Generation of Diverse Knowledge Form

Deepa Tilwani,
Phd Candidate
University of South Carolina

Introduction and Motivation (Part 1)

Progress in Language Modelling

Symbolic Era

Pre - 1990

- Rule Based
- Expert Systems
- Limited Generalization

Statistical Era

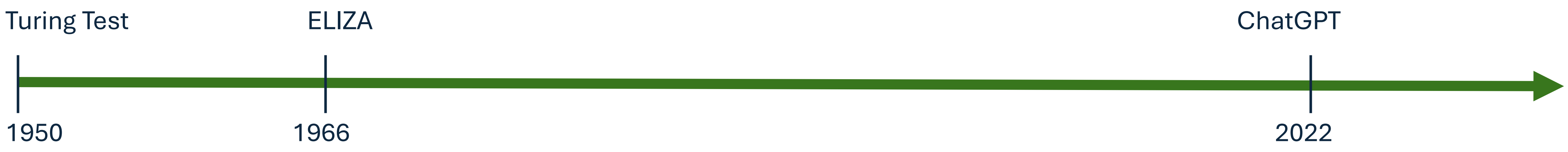
1990 - 2006

- Data-driven Approaches
- Probabilistic Models

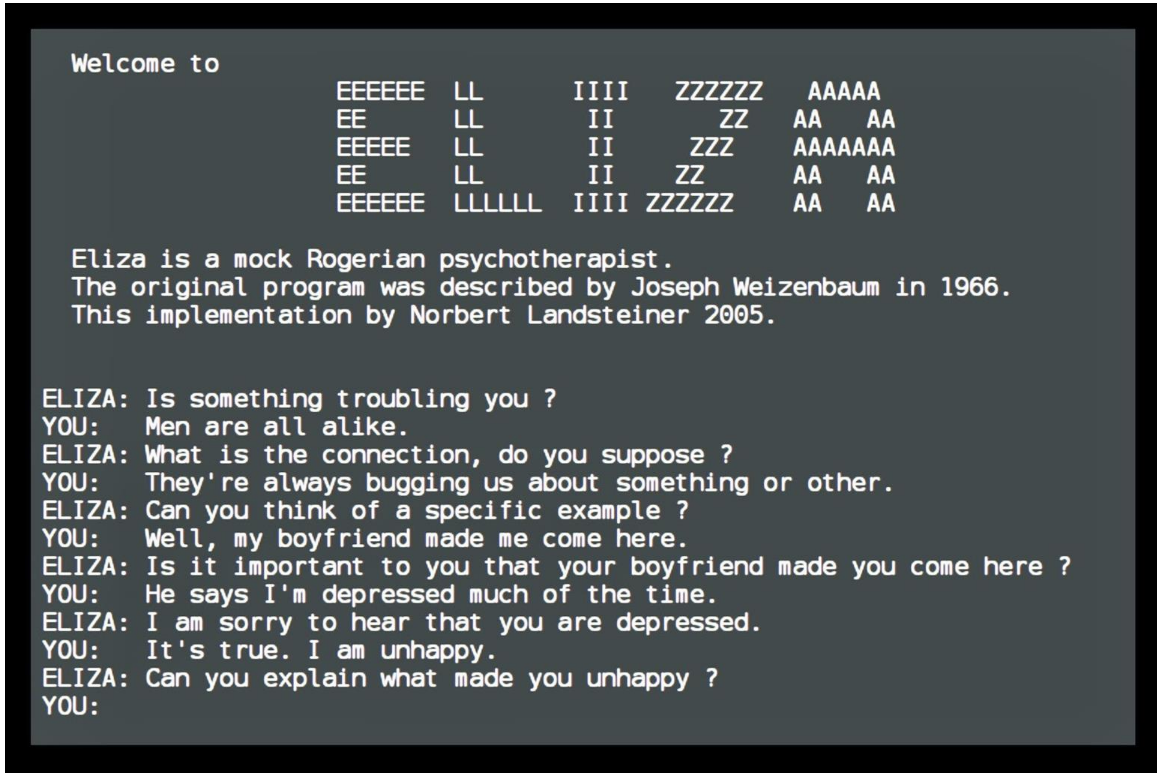
Scale Era

2006 onwards

- Deep Learning and neural nets
- General Purpose LMs
- Massive Datasets and Computation



ELIZA (1966) : THE FIRST CHATBOT



Early NLP program developed by Joseph Weizenbaum at MIT. Created illusion of a conversation by rephrasing user statements as questions using pattern matching and substitution methodology. One of the first programs capable of attempting the Turing test.

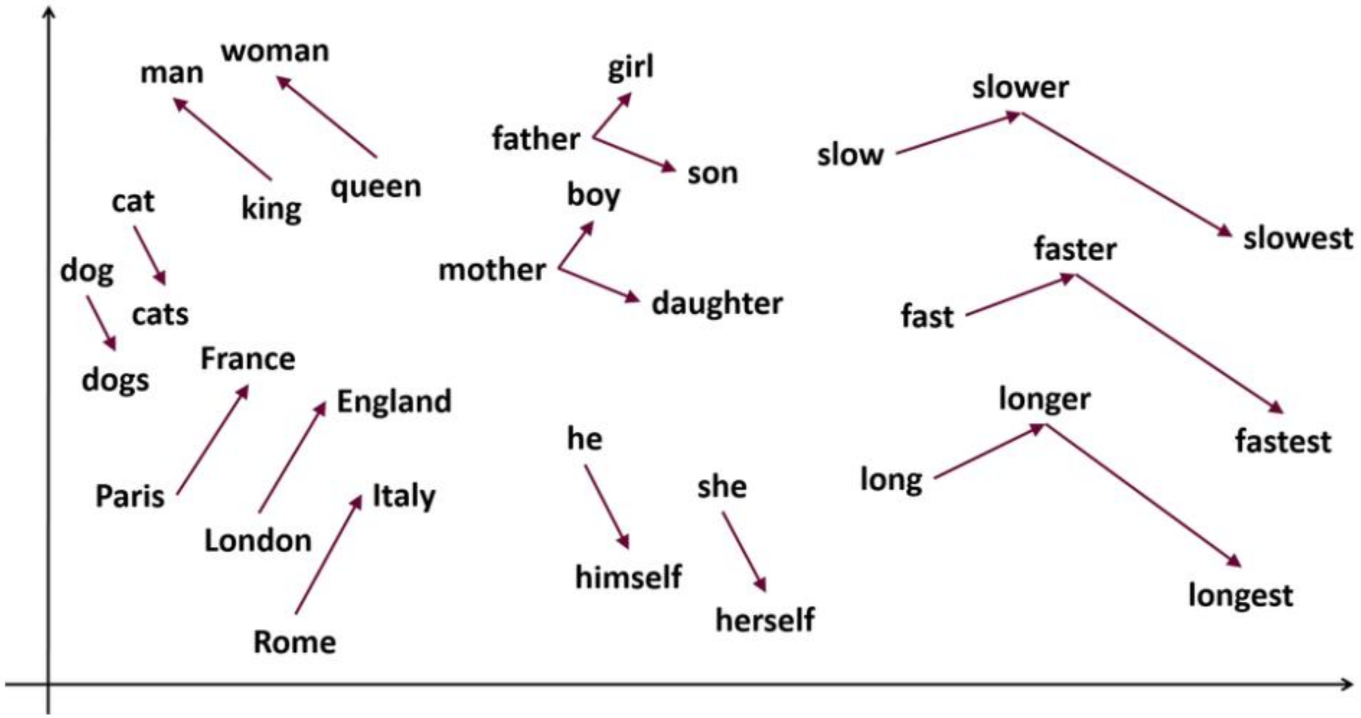
Try it out at https://web.njit.edu/~ronkowit/eliza.html

The LLM Era – How they work?

Word Embeddings

Represent each word using a “vector” of numbers.

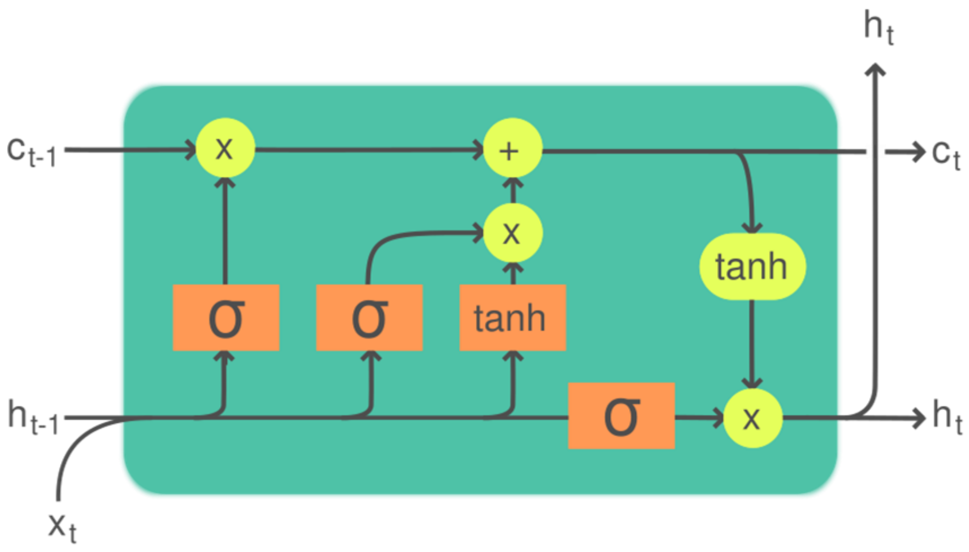
- Converts a “discrete” representation to “continuous”.
- Many benefits:
  - More “fine-grained” representations o words.
  - Useful computations such as cosine and Euclidean distance.
  - Visualization and mapping of words onto a semantic space.
  - Can be learnt in self-supervised manner from a large corpus.
- Examples:
  - Word2Vec (2013), GloVe, BERT, ELMo

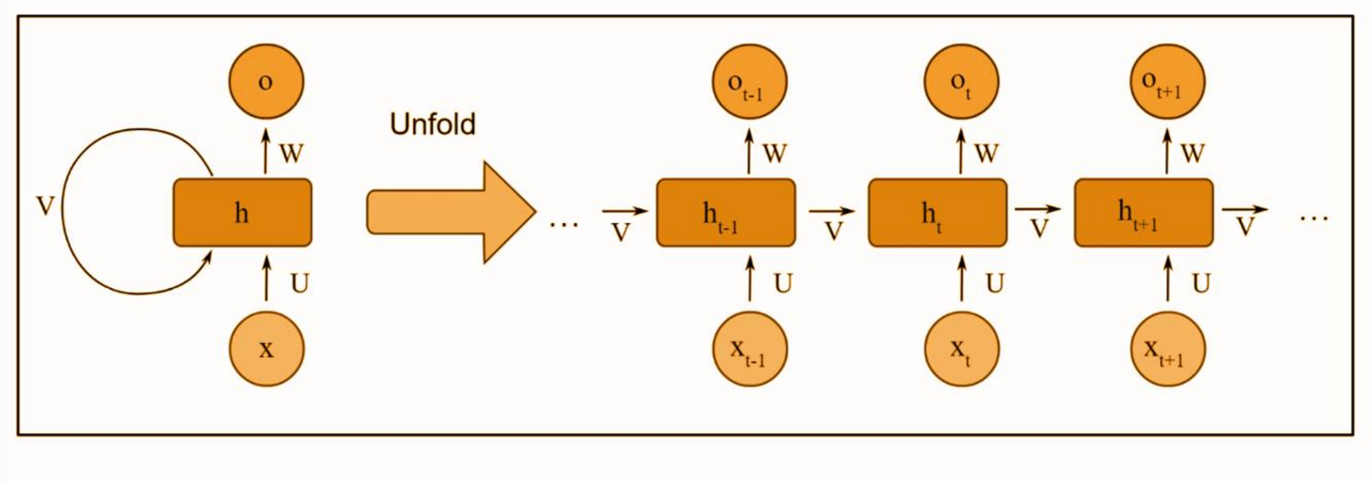


Seq2Seq Models

Recurrent Neural Networks (RNNs)

- Long Short-Term Memory Networks (LSTMs)
- Capture dependencies between input tokens
- Gates control the flow of information

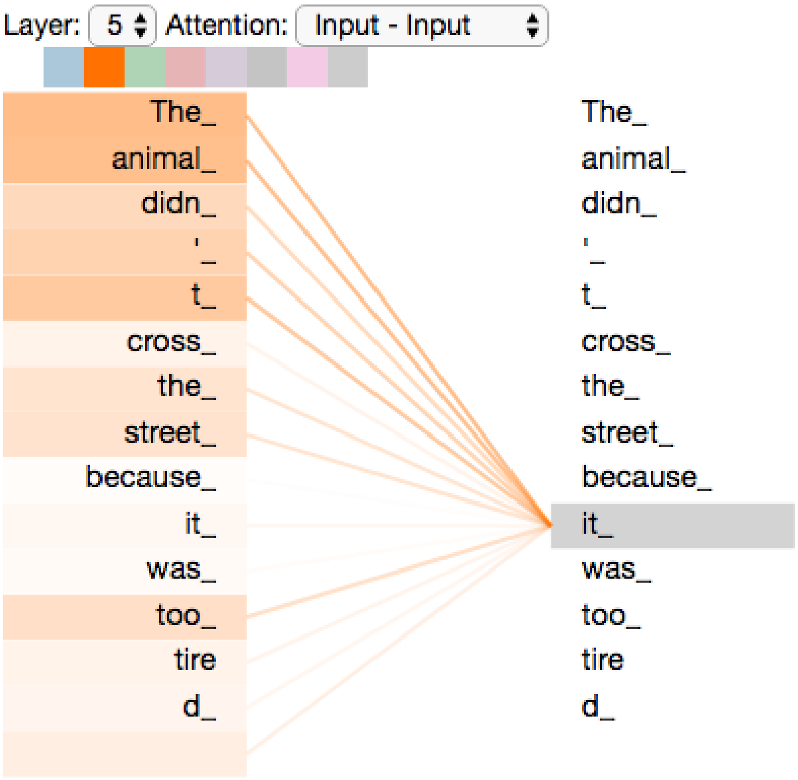




A simple RNN shown unrolled in time. Network layers are recalculated for each time step, while weights U, V and W are shared across all

- The inputs to each unit consists of the current input xt, previous hidden state h_{t-1}, and previous context c_{t-1}
- The outputs are a new hidden state h_t and an updated context c_t.

Transformers



- Allows to "focus attention" on particular aspects of the input while generating the output.
- Done by using a set of parameters, called "weights," that determine how much attention should be paid to each input token at each time step.
- These weights are computed using a combination of the input and the current hidden state of the model.

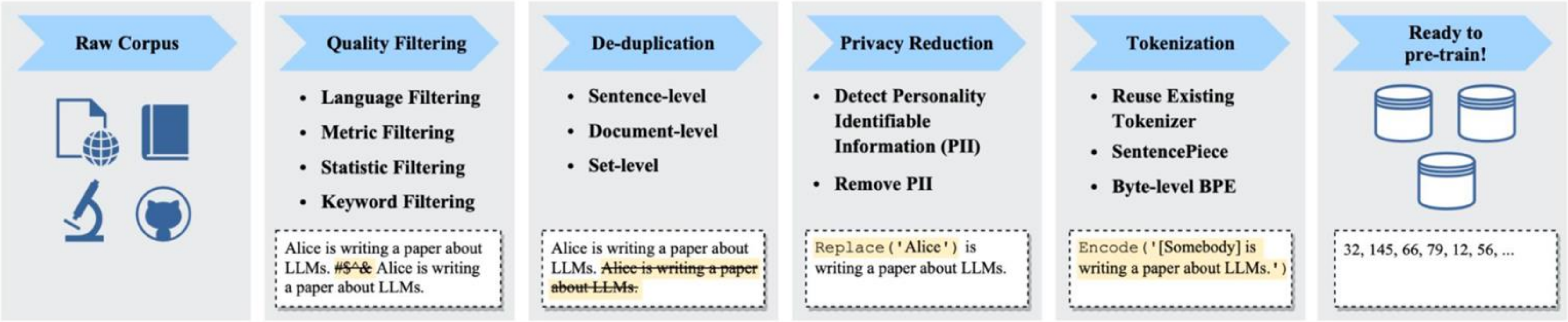
In encoding the word "it", one attention head is focusing most on "the animal", while another is focusing on "tired". The model's representation of the word "it" thus bakes in some of the representation of both "animal" and "tired".

https://jalammar.github.io/illustrated-transformer/

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Pre-Training: Data Preparation

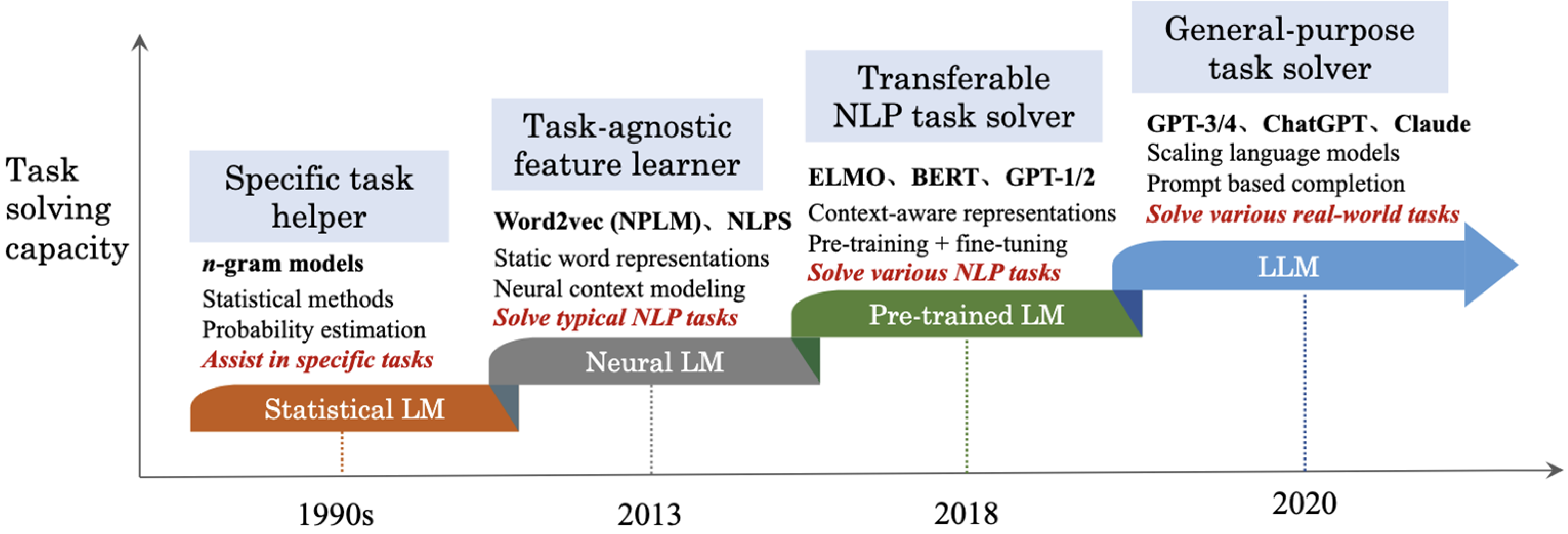
A typical data preparation pipeline for pre-training LLMs:



W. Zhao et al. A Survey of Large Language Models. 2023.

What LLMs Can do?

Evolution of LMs from Perspective of Task-Solving Capacity

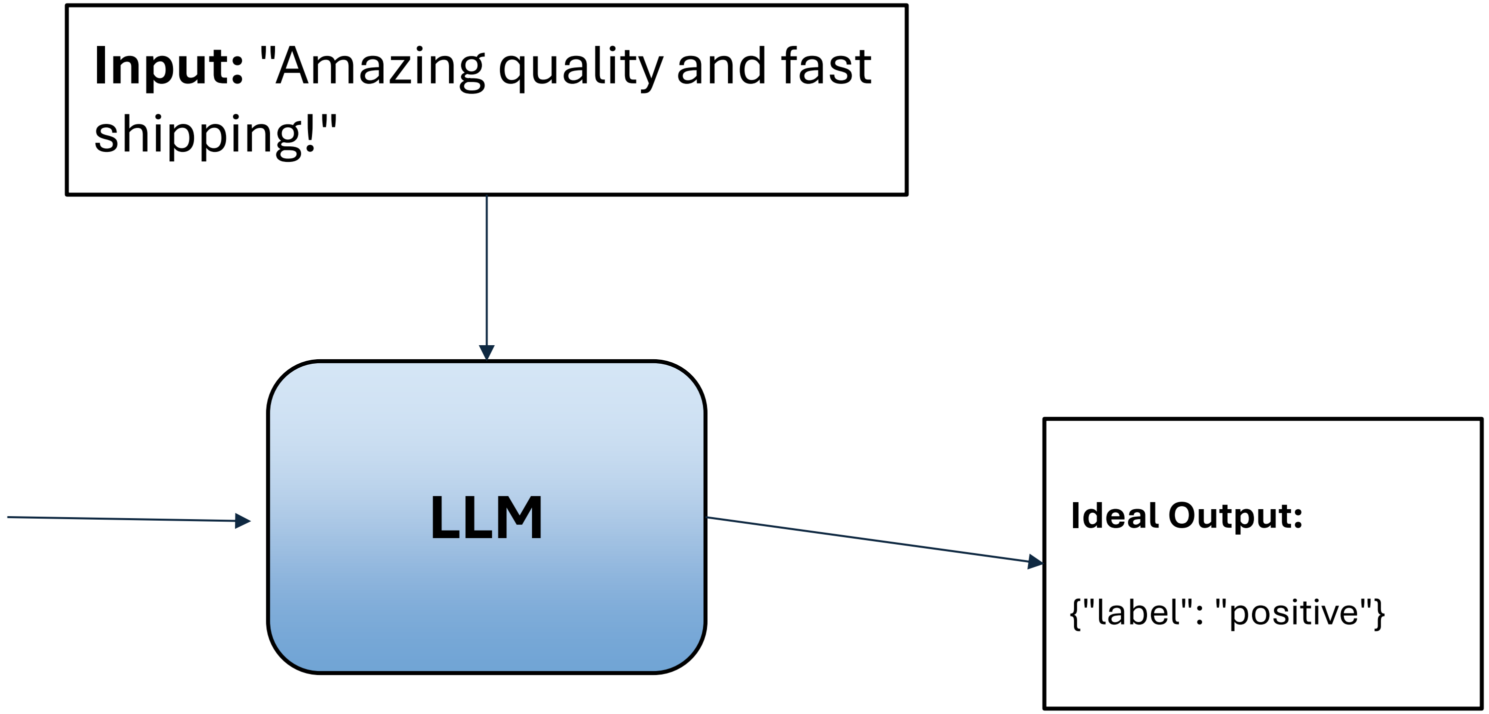


W. Zhao et al. A Survey of Large Language Models. 2023.

Few-Shot Prompting

Instruction:
Classify the sentiment of the given text as either **positive** or **negative** based on the examples provided.
Few shots examples:

"Great product, 10/10":
{"label": "positive"}
"Didn't work very well":
{"label": "negative"}
"Super helpful, worth it":
{"label": "positive"}



T. Brown et al. Language Models are Few-Shot Learners. NeurIPS 2020.

Chain-of-Thought Prompting

Instruction:
Classify the sentiment of the given text as either **positive** or **negative**. Follow a step-by-step reasoning process to determine the sentiment.

Examples:

1. **Input:** "Great product, 10/10"
   Reasoning: The phrase "Great product" expresses strong approval, and "10/10" indicates a perfect rating, showing high satisfaction.
   Output: {"label": "positive"}
2. **Input:** "Didn't work very well"
   Reasoning: The phrase "Didn't work" suggests malfunction or failure, and "very well" implies that it performed below expectations. This conveys dissatisfaction.
   Output: {"label": "negative"}
3. **Input:** "Super helpful, worth it"
   Reasoning: "Super helpful" indicates a high level of usefulness, and "worth it" suggests that the person finds the product valuable. This implies strong satisfaction.
   Output: {"label": "positive"

Input: "Wow! This is fantastic quality and fast shipping!"

LLM

Reasoning:

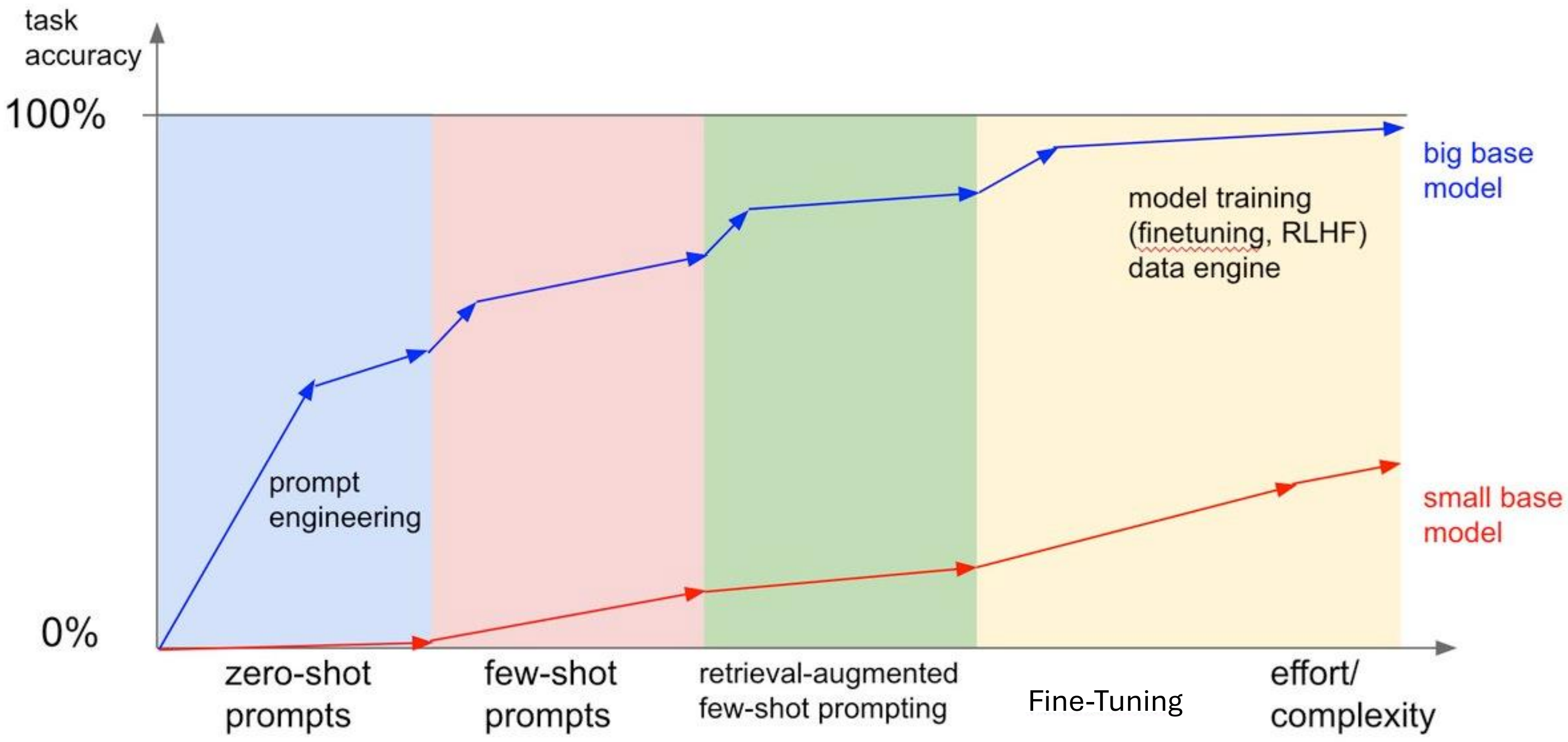
- "Wow!" expresses excitement and enthusiasm, indicating a positive reaction.
- "This is fantastic quality" suggests high satisfaction with the product's quality.
- "Fast shipping" is another positive aspect, showing appreciation for timely delivery.
- All elements in the sentence convey strong positivity.

Output: {"label":"positive"}

J. Wei et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. NeurIPS 2022.

From Prompting to Fine-Tuning

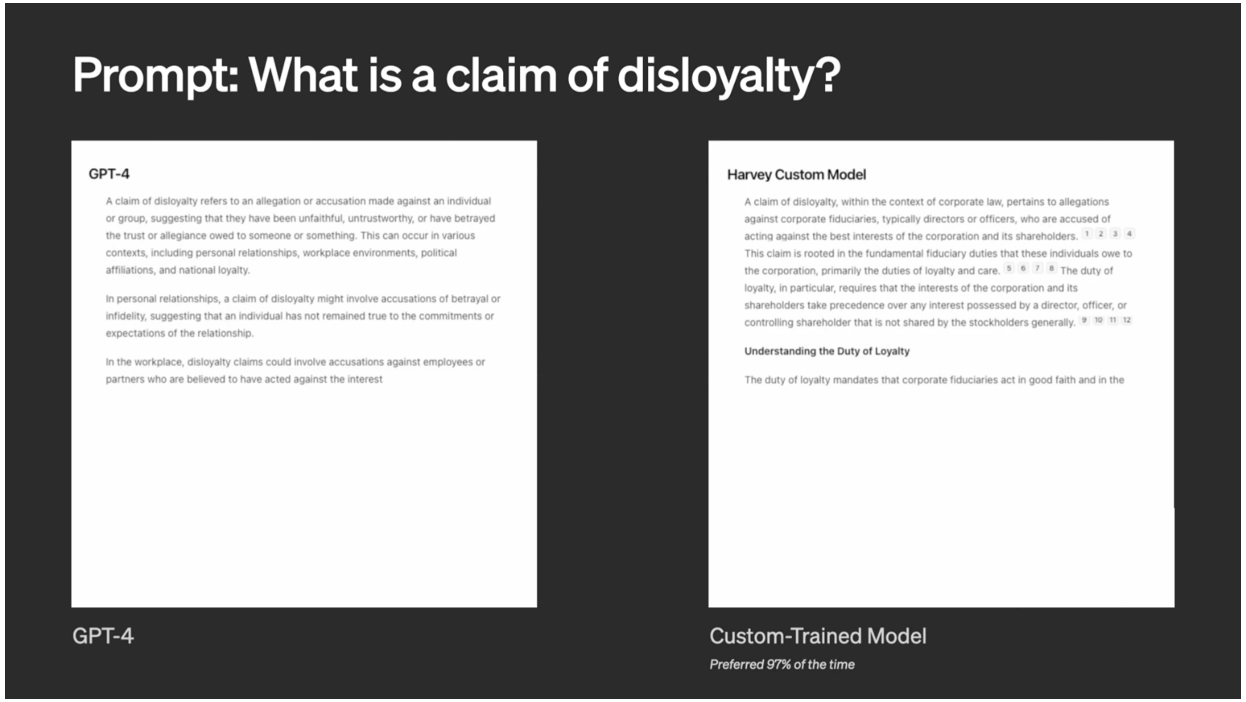
Unlike prompting, fine-tuning actually changes the model under the hood, giving better domain- or task-specific performance.



https://x.com/karpathy/status/1655994367033884672

Custom Trained Model in Law: Harvey AI

- Startup building a custom-trained case law model for drafting documents, answering questions about complex litigation scenarios, and identifying material discrepancies between hundreds of contracts.
- Added 10 billion tokens worth of data to power the model, starting with case law from
- Delaware, and then expanding to include all of U.S. case law.



Open AI Customer Stories: Harvey. April 2024.

Why the need for Trustworthiness in Generative AI?

Unreliable Reasoning Even On Simple Tasks

Hi name is deepa and how many e's are in my name.

Hi Deepa! Your name has **two** "e"s in it. 😊

Easy reasoning, Sure!

9.11 and 9.9, which one is larger?

9.11 is larger than 9.9.

Got confused ??

Probably due to tokenization!

Generated by gpt-4o’s tokenizer.

Try it out at:

https://tiktokenizer.vercel.app/

The Story of a Lawyer Who Employed ChatGPT ... trust issues remain

The New York Times

Artificial Intelligence › | A.I. Faces Quiz | How the A.I. Race Began | Key Figures in the Field | One Year of C

Here's What Happens When Your Lawyer Uses ChatGPT

A lawyer representing a man who sued an airline relied on artificial intelligence to help prepare a court filing. It did not go well.

Share full article | 1.1K

A lawyer, representing a client against an airline, turned to AI assistance for drafting legal documents. The results were less than ideal. https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html

The New York Times

ChatGPT Lawyers Are Ordered to Consider Seeking Forgiveness

Steven A. Schwartz and Peter LoDuca must pay a fine and send letters to judges named in a brief filled with fiction, a judge ordered.

Share full article

Legal Consequences for Attorneys Using ChatGPT

https://www.nytimes.com/2023/06/22/nyregion/lawyers-chatgpt-schwartz-loduca.html

The New York Times

Intelligence › | A.I. Faces Quiz | How the A.I. Race Began | Key Figures in the Field | One Year of

The ChatGPT Lawyer Explains Himself

In a cringe-inducing court hearing, a lawyer who relied on A.I. to craft a motion full of made-up case law said he "did not comprehend" that the chat bot could lead him astray.

Share full article | 267

Lawyer Acknowledges AI Misuse in Court: During court session, an attorney admitted excessively relying on AI, resulting in a legal motion filled with artificial legal references.
https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html

Neuro Symbolic Legal AI

Chapter 1

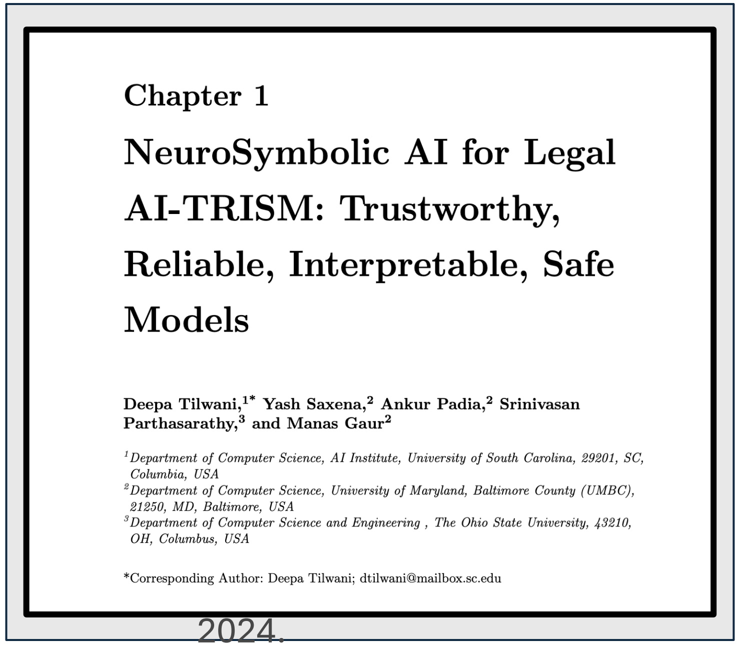
NeuroSymbolic AI for Legal AI-TRISM: Trustworthy, Reliable, Interpretable, Safe Models

Deepa Tilwani,[1*] Yash Saxena,[2] Ankur Padia,[2] Srinivasan Parthasarathy,[3] and Manas Gaur[2]

[1] *Department of Computer Science, AI Institute, University of South Carolina, 29201, SC, Columbia, USA*
[2] *Department of Computer Science, University of Maryland, Baltimore County (UMBC), 21250, MD, Baltimore, USA*
[3] *Department of Computer Science and Engineering , The Ohio State University, 43210, OH, Columbus, USA*

*Corresponding Author: Deepa Tilwani; dtilwani@mailbox.sc.edu

Neuro Symbolic Legal AI


Chapter 1

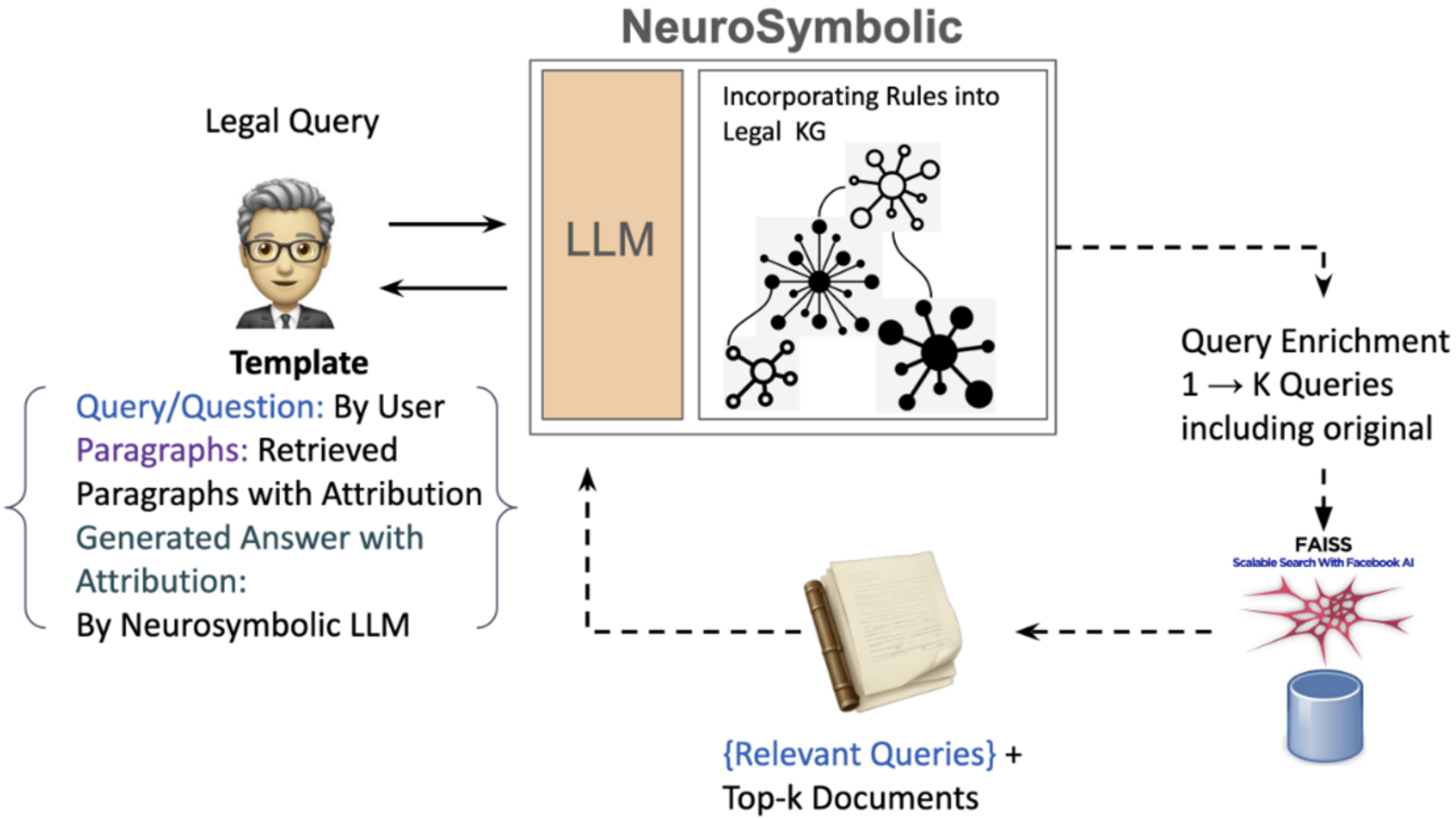
NeuroSymbolic AI for Legal AI-TRISM: Trustworthy, Reliable, Interpretable, Safe Models

Deepa Tilwani,[1*] Yash Saxena,[2] Ankur Padia,[2] Srinivasan Parthasarathy,[3] and Manas Gaur[2]

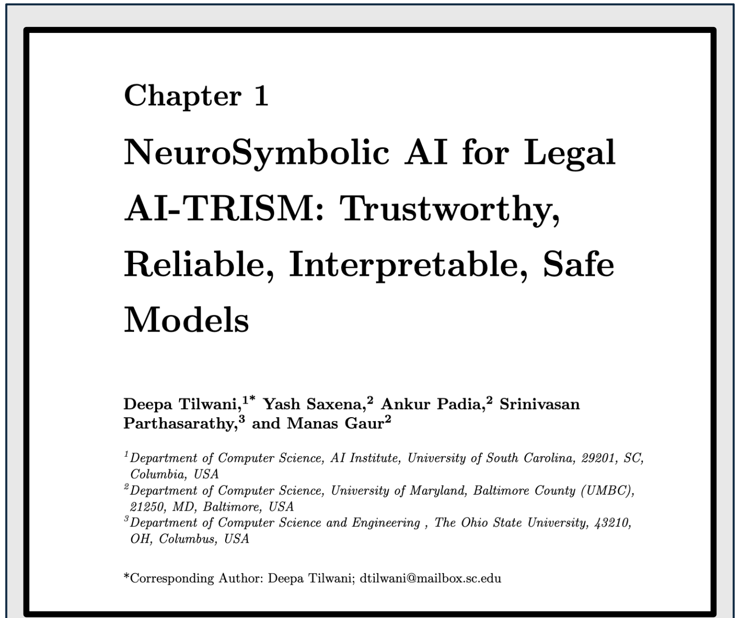
[1]Department of Computer Science, AI Institute, University of South Carolina, 29201, SC, Columbia, USA
[2]Department of Computer Science, University of Maryland, Baltimore County (UMBC), 21250, MD, Baltimore, USA
[3]Department of Computer Science and Engineering , The Ohio State University, 43210, OH, Columbus, USA

*Corresponding Author: Deepa Tilwani; dtilwani@mailbox.sc.edu

2024.



Neuro Symbolic Legal AI


Chapter 1

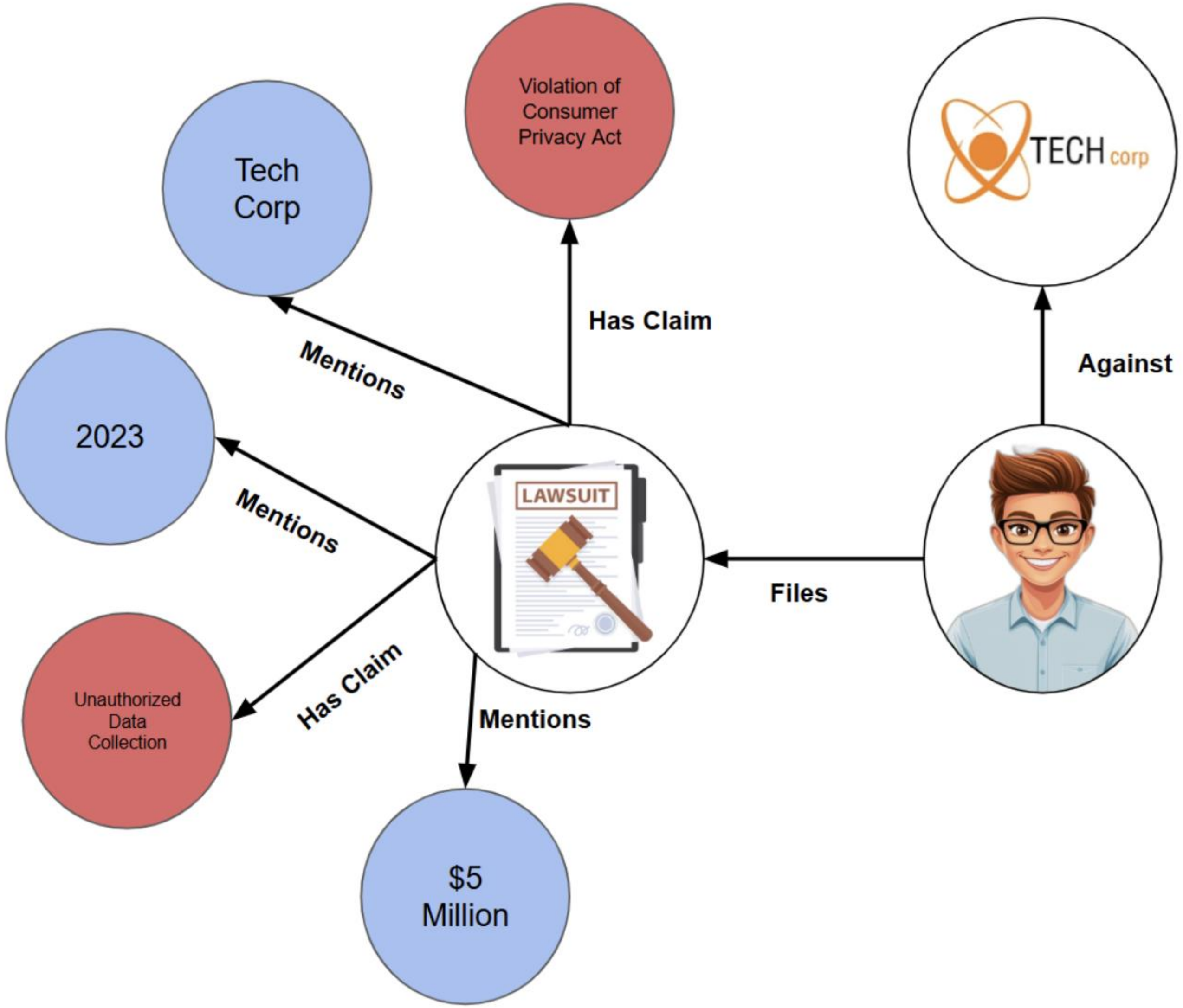
NeuroSymbolic AI for Legal AI-TRISM: Trustworthy, Reliable, Interpretable, Safe Models

Deepa Tilwani,[1*] Yash Saxena,[2] Ankur Padia,[2] Srinivasan Parthasarathy,[3] and Manas Gaur[2]

[1]Department of Computer Science, AI Institute, University of South Carolina, 29201, SC, Columbia, USA
[2]Department of Computer Science, University of Maryland, Baltimore County (UMBC), 21250, MD, Baltimore, USA
[3]Department of Computer Science and Engineering , The Ohio State University, 43210, OH, Columbus, USA

*Corresponding Author: Deepa Tilwani; dtilwani@mailbox.sc.edu

2024.



More Challenges for the Generative AI

- Inconsistency in Prompts for Completeness in Outcomes
- Nearly Impossible to Explain or Reason Generative Answers
- Prompt Injections can leak data
- Context Windows are and will remain limited
- Bias in Large Language Models that Supervised Learning cannot reduce
- Reliability Issues: Different Large Language Models Yield Different Outcomes

Grounding (Part 2)

Grounding

Grounding is defined as ensuring **every claim in an LLM response** generate **verifiable and well-grounded** responses to any prompt, relying solely on information from **a user-specified knowledge base.**

Grounded means that every claim in the response is *attributable to a document in the knowledge base

Knowledge base may be a private corpus, a public domain, entire Web
E.g., a healthcare customer may specify a set of journals they trust

Verifiably-grounded means that every claim is backed by an appropriate citation

Two Core Approaches to Grounded AI

Grounded Generation – Enhancing AI with Verified Knowledge

Method:

- Retrieve relevant facts from a trusted **knowledge base**.
- Augment LLM prompts with **verified context** before generating responses.
- Intrinsic phenomena

Grounding Verification – Ensuring AI's Responses Are Factually Correct

Method:

- **Cross-check AI-generated claims** with authoritative sources.
- **Generate citations** to improve transparency and accountability.
- Apply **fact-checking models** to filter unverified claims.
- Extrinsic phenomena

Why Grounded Generation ?

*“**Grounded generation retrieves latest clinical guidelines** and provides an evidence-based response”*

Input

I feel like the only reason I haven't already found the courage to do it is my two younger siblings. I'm sixteen, and they are twelve and ten respectively. I know that my parents would be heartbroken, but that would only be two people. My "friends" would be over it in a few weeks anyway. I feel like my brother and sister would be too emotionally damaged or traumatised if I did it, and I don't want to cause any more pain to anyone anymore.

Ground Truth Explanation: emotionally damaged, traumatised

LLAMA

I'm really sorry to hear that you're feeling this way, suicidal thoughts are serious, and it's crucial. Remember, your family care about you and want to help you through difficult times.

Not grounded but a generic answer!

Why Grounding Verification ?

INPUT:
What is the target blood pressure for men?

It should first verify ***who the intended audience is*** *before ensuring factual accuracy.*

National Institutes of Health (.gov)
https://www.ncbi.nlm.nih.gov › articles › PMC6813156

Ideal Target Blood Pressure in Hypertension - PMC

by S Park · 2019 · Cited by 34 — Based on these results, the ACC/AHA arterial hypertension guideline recommended a target **SBP <130 mmHg** for elderly hypertensive patients. However,...

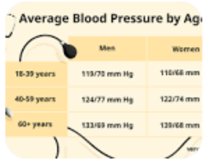
Not according to 2017 guidelines

What is the target blood pressure for men?

All Images Videos Shopping Forums Web News ⋮ More

AI Overview

According to the American Heart Association, the target blood pressure for men (and most adults) is considered to be a systolic pressure of less than 120 mmHg and a diastolic pressure of less than 80 mmHg, meaning a "normal" blood pressure reading is below 120/80 mmHg.



Key points about blood pressure targets for men:

Normal range:
Less than 120/80 mmHg

Elevated range:
Systolic pressure between 120-129 mmHg with a diastolic pressure less than 80 mmHg

High blood pressure (Stage 1):
Systolic pressure between 130-139 mmHg or diastolic pressure between 80-89 mmHg

High blood pressure (Stage 2):
Systolic pressure 140 mmHg or higher or diastolic pressure 90 mmHg or higher

Important note: While this is the general guideline, it's crucial to consult with your doctor to determine the best target blood pressure for your individual health needs and circumstances, including age and any existing medical conditions.

This is for informational purposes only. For medical advice or diagnosis, consult a professional. Generative AI is experimental.

Types of Grounding in AI & LLMs

1) Symbolic Grounding

AI must **retrieve**, **recognize**, and **structure information correctly** before using it (i.e. LLMs should understand and link symbols like words, phrases, numbers to their real-world meanings)

a) Retrieval-Augmented Generation Based Grounding

- **Method:** AI **retrieves external documents** before responding.
- **Example:**
  - A **QA system fetching Wikipedia articles** before answering a historical question.
  - **Chatbot retrieving product manuals** before explaining a feature.

a) Knowledge Graph-Based Grounding

- **Method:** AI **structures information in graphs** to improve contextual understanding.
- **Example:**
  - A **search engine linking related topics** (e.g., AI connects "COVID-19" with "vaccines" and "pandemics").
  - **Legal NLP models linking case laws, statutes, and judicial precedents** to provide structured responses.

2) Functional Grounding

LLMs should **reason, verify, and adapt responses based on context** (i.e **apply it correctly** in context)

a) Attribution-Based Grounding

- **Method:** AI **justifies responses** with references and citations.
- **Example:**
  - A **fact-checking AI cross-referencing news claims** with verified sources before publishing.
  - AI writing research papers by **citing scientific studies** instead of generating unsupported claims.

a) Interactive & Reinforcement-Based Grounding

- **Method:** AI **learns from real-time feedback** and **improves responses over time**.
- **Example:**
  - **Customer support chatbots adapt based on user corrections** (e.g., learning new slang or product updates).
  - **Language models are fine-tuned through user interactions** to generate more accurate, personalized responses.

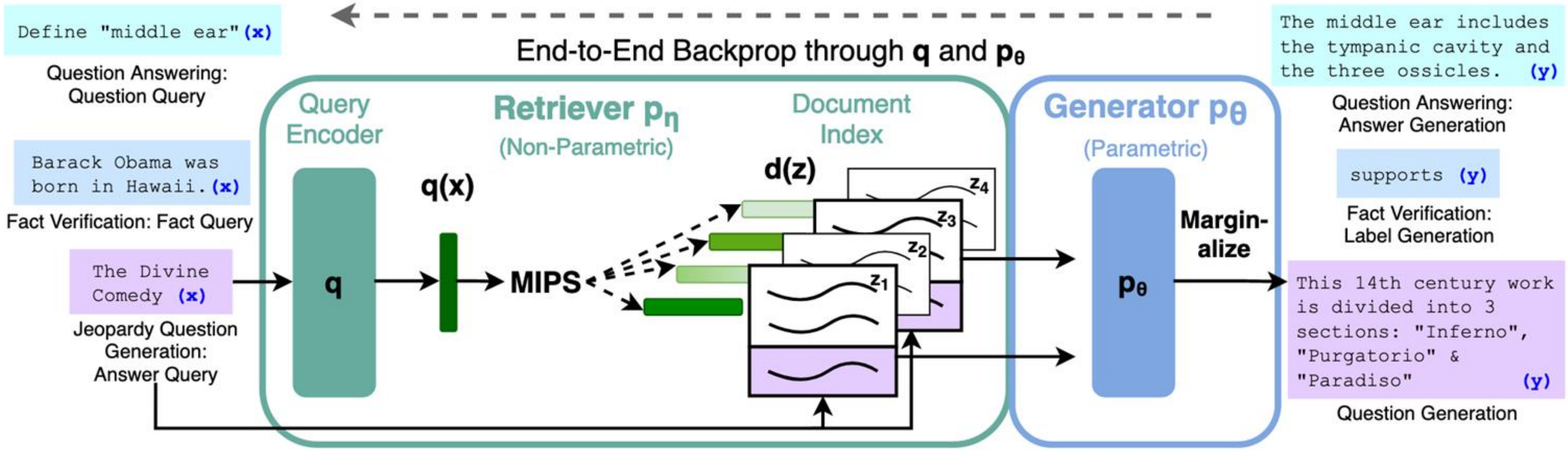
Symbolic Grounding

1. Retrieval-Augmented Generation (RAG)

LLMs lack knowledge beyond their training date, and frequent model updates are impractical.

Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." *Advances in Neural Information Processing Systems* 33 (2020): 9459-9474.

Idea: Enhance LLMs with a retrieval system!



Advantages

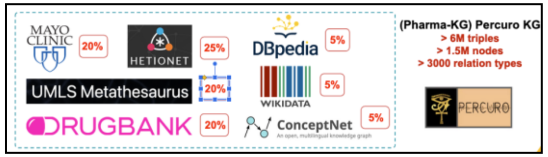
Accessible and Affordable

Safe

Cost-Effective

Custom Train

Domain Knowledge



Fact-Checking

Continuous Update

Easier to Customize

Symbolic Retrieval based Grounding

Source Attribution: **Retrieve**, **recognize**, and **attribute**

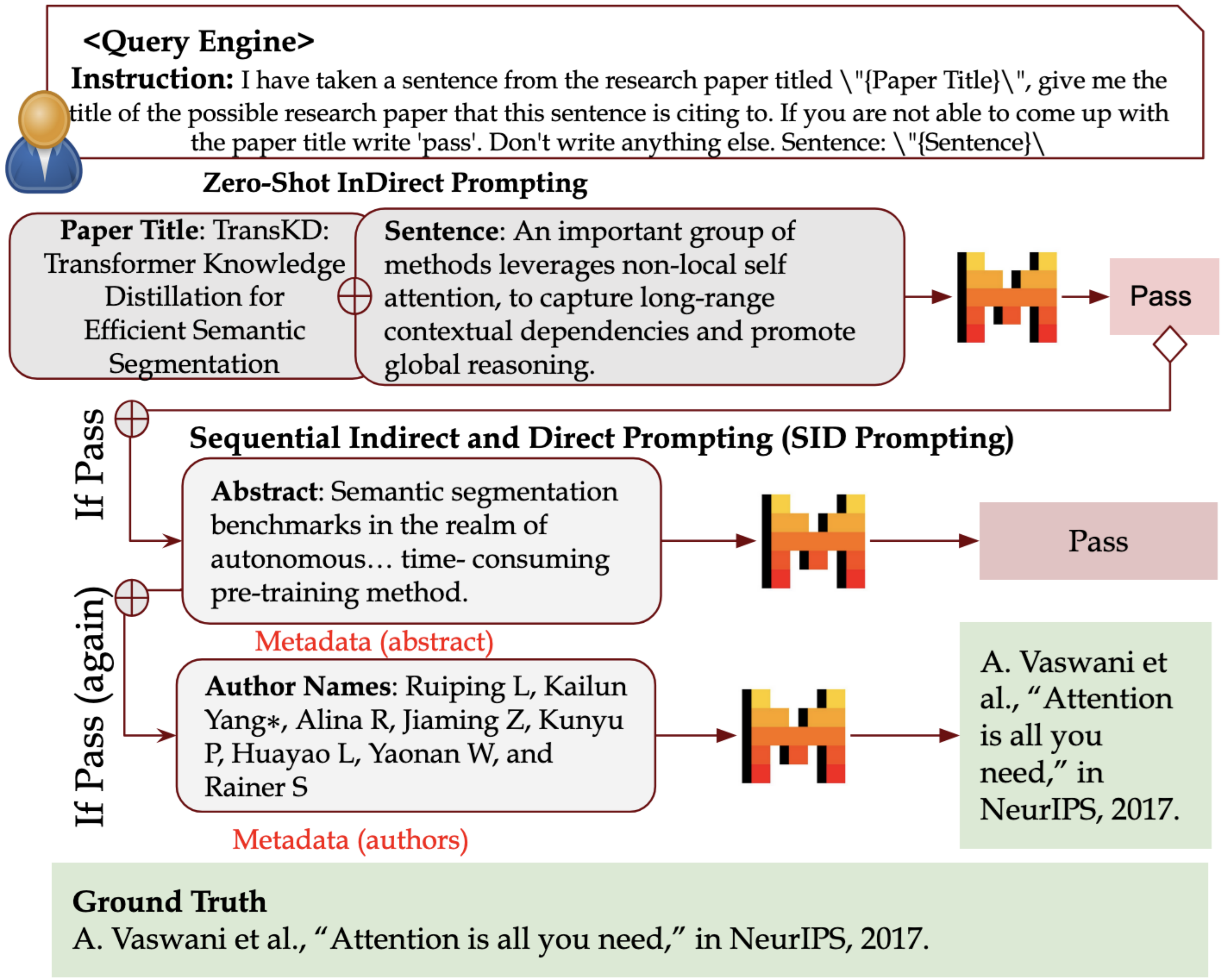
- RAG enabled system
- Domain-Specific Training
- Enhanced Accuracy and Relevance
- Customization for Business Needs
- Business Alignment

An Evaluation study of Citation Generation on Recent LLMs

Grounded and targeted for generating citations with structured metadata

Image: Tilwani, Deepa, et al. "REASONS: A benchmark for REtrieval and Automated citationS Of scieNtific Sentences using Public and Proprietary LLMs." *arXiv preprint arXiv:2405.02228* (2024).

How we do Symbolic Retrieval based Grounding ?



But Few Limitations of RAG..

Needs Existing Database

Latency Issues

Hallucination still exists

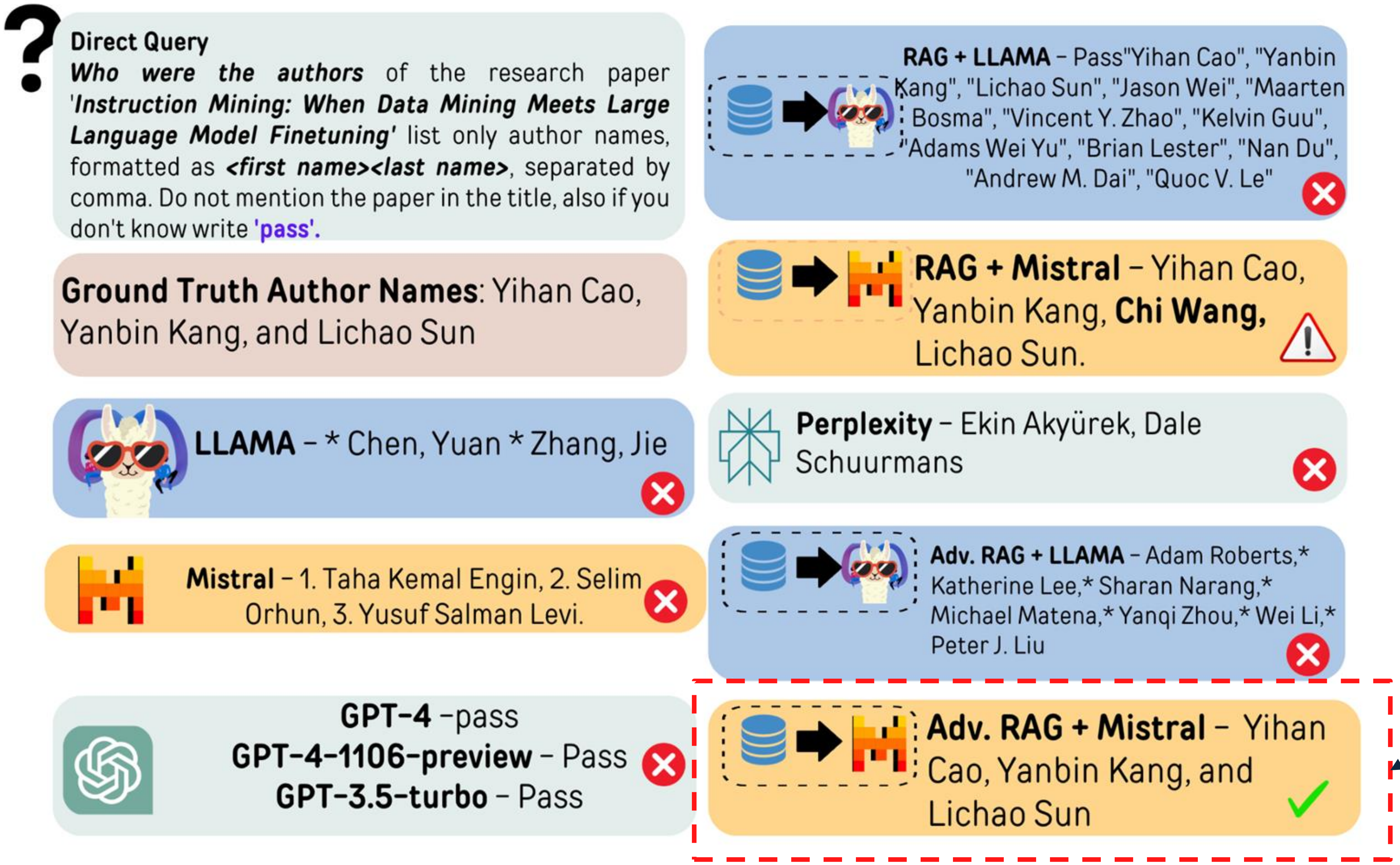
Context Length Limitation

Dependent on Semantic Search

At scale, sensitive to choices of:

1) Chunking Strategy,
2) Embedding Model, and
3) Generation Model.

Flaws in RAG from REASONS Dataset



Only Adv. RAG was able to correctly generate author names

Latency Issues

Domain	OpenAI	M	L	D	RM	RL	P	AdvRAG(L)	AdvRAG(M)
AI	34:25	26:03	11:10	34:11	74:49	73:09	34:31	156:24	163:28
CV	47:45	18:35	19:24	50:22	189:20	198:45	42:05	259:32	302:14
Cryptography	03:50	02:18	04:59	32:21	83:28	89:21	13:23	190:19	194:25
Graphics	07:08	08:55	06:08	58:43	108:08	127:48	16:52	214:25	227:23
HCI	03:01	01:10	00:42	21:56	48:32	50:51	02:47	95:56	98:44
IR	20:31	11:40	06:52	33:34	91:30	99:43	19:50	193:37	202:23
NLP	28:26	11:42	05:09	47:24	91:07	88:40	13:06	175:58	156:49

2. Knowledge Graphs (KG) Based Grounding

1. Machine-readable structured representation of knowledge
2. Consisting of <u>entities, entity types, and relationships</u> in various forms (e.g., ontologies, lexicons, labeled property graphs and RDFs).

Subject	Predicate	Object
World War I	fought_with	Poisonous Gas

WIKIDATA

empathi

ConceptNet

WORDNET

"KG-based grounding structures information in graphs, linking concepts to improve AI systems' ability to ***retrieve and generate meaningful responses****."*

Speer et al. AAAI'17
Vrandečić et al. ACM Comm'14
Gaur et al. ICSC'19
Miller, ACM Comm'95

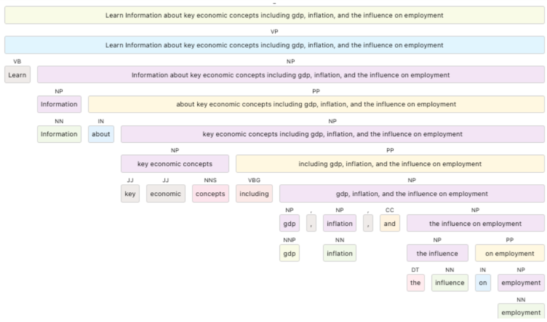
Illustration of ISEEQ

Query

Title: Economy and Employment Statistics
Description: Learn Information about key economic concepts including gdp, inflation, and the influence on employment

Constituency Parsing

Information + { economy, employment statistics, employment, influence employment, inflation influence employment, gdp, gdp influence employment, key economic concepts}

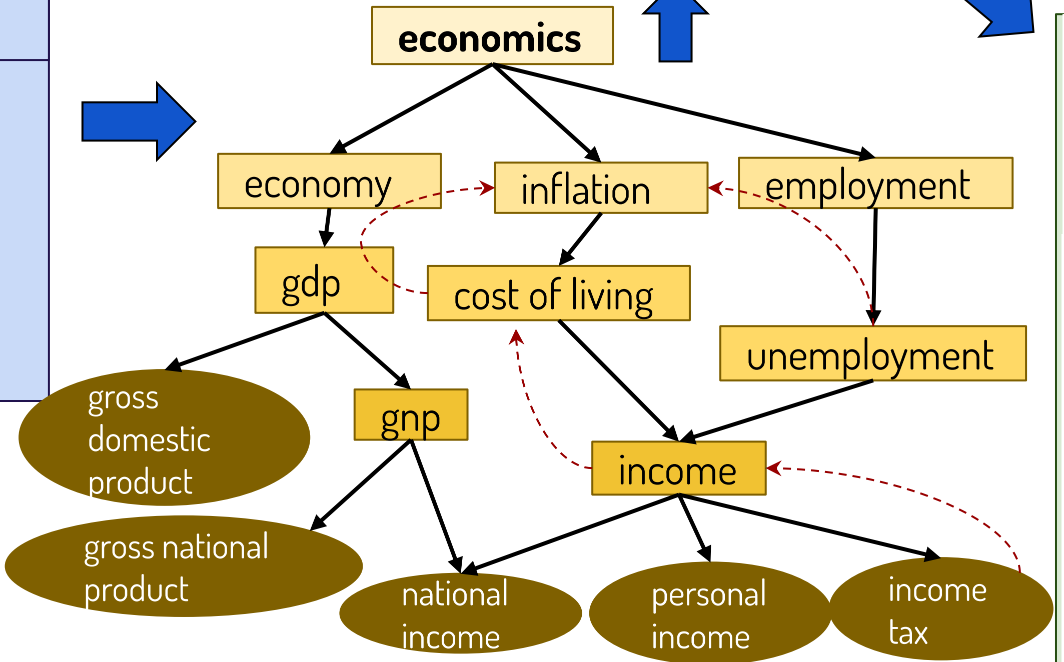


ISQ by Generative Adversarial Reinforcement Learning

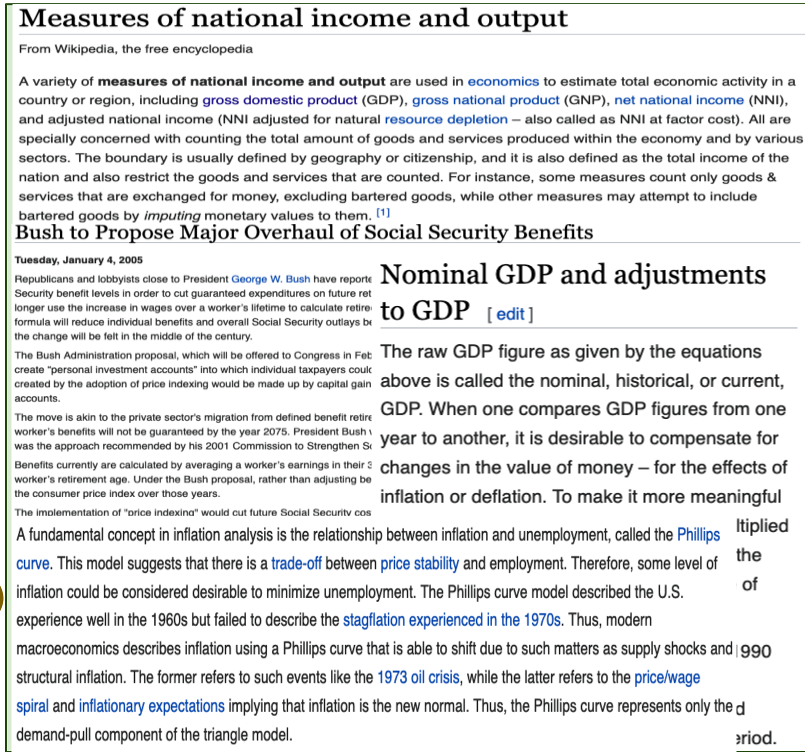
1. What is gross_domestic_product?
2. What is the measure of gross_domestic product?
3. **What is the reason nation income relations gross_domestic_product? gross_domestic_product?**
4. What is the influence of inflation to gross_domestic_product?
5. **What is the meaning of unemployment in inflation?**
6. **What is the influence of inflation on cost_of_living?**

Sentence BERT Encoder

Sentence BERT Encoder

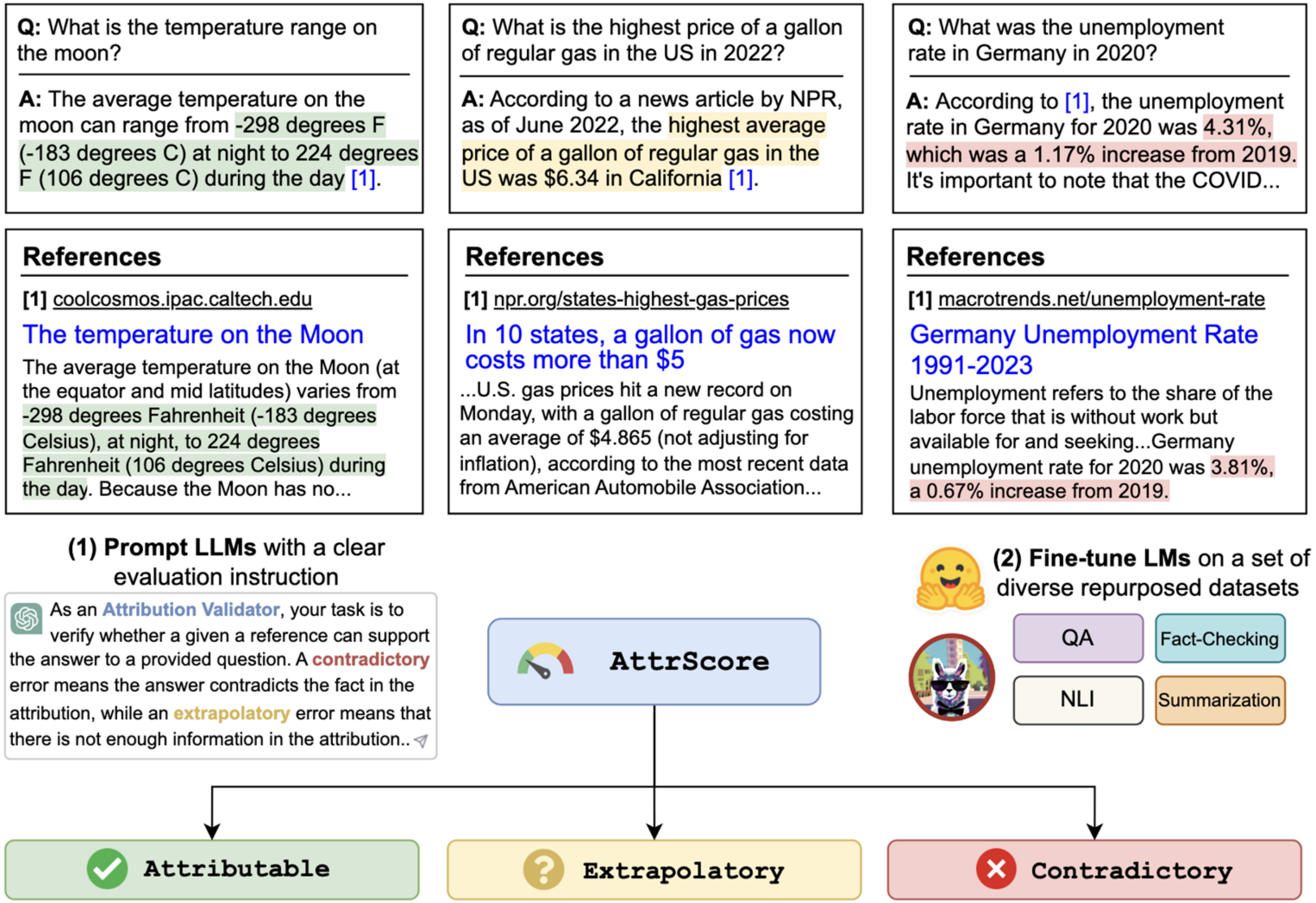


ConceptNet Graph for Semantic Query Expansion



Functional Grounding

Attribution-Based Functional Grounding



Evaluating attribution and identifying specific types of errors with AttrScore. We explore two approaches in AttrScore: (1) prompting LLMs, and (2) fine-tuning LMs on simulated and repurposed datasets from related tasks

Helps to work on :

- Factual verification
- Reducing hallucinations
- External knowledge integration
- Error categorization

A ***fact-checking approach for cross-referencing news claims*** *with verified sources before publishing.*

Yue, Xiang et al. “Automatic Evaluation of Attribution by Large Language Models.” *EMNLP* (2023).

Reinforcement Learning for Grounding

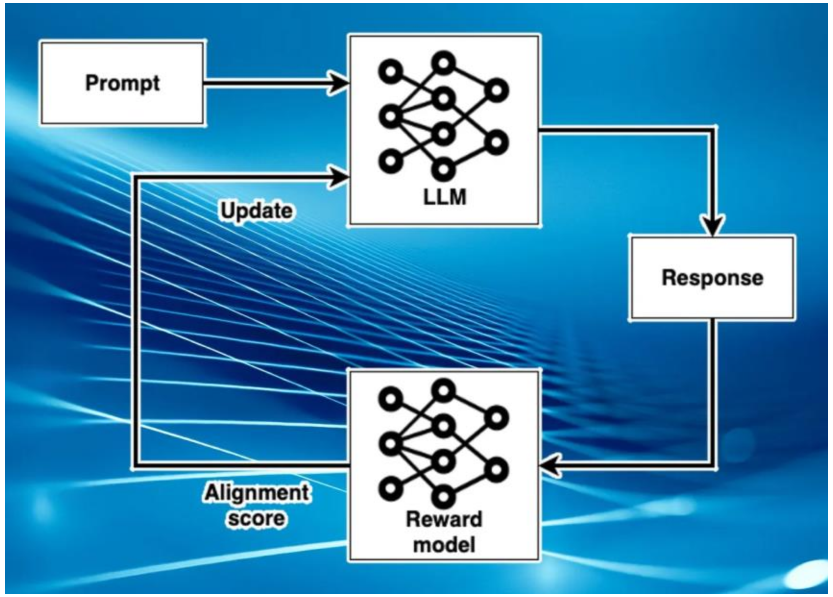
What constitutes a good response for a query and context is quite nuanced?

Idea: Capture this using a **reward model** that scores <query, context, response> on the appropriateness of the response. The model may be trained on a dataset **preferences between response pairs**

We can then use **reinforcement learning** to tune the to maximize reward while staying within a bounded KL-divergence from the initial model.

References:

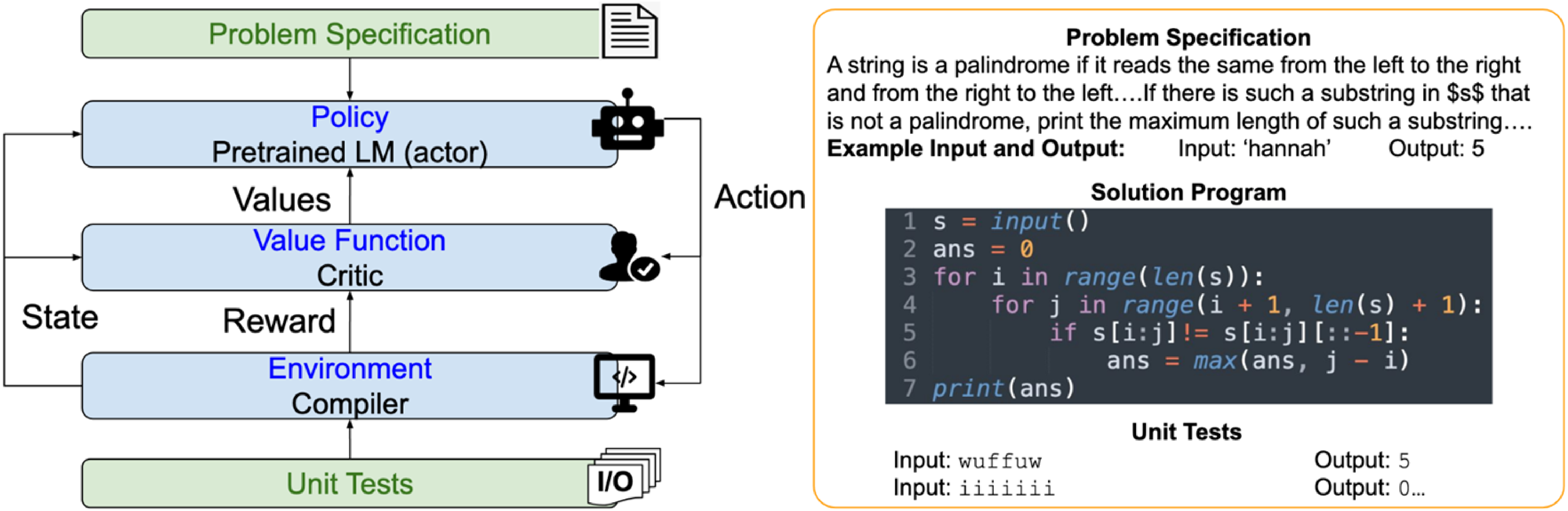
- Teaching language models to support answers with verified quotes
- Fine tuning language models for factuality



Src: https://bdtechtalks.com/2023/01/16/what-is-rlhf/

Interactive & Reinforcement-Based Grounding

Interactive & Reinforcement-Based Grounding *ensures that LLMs do not just generate blindly but engage in a feedback-*



The code-generating language models as an actor network, and introduce a critic network that is trained to predict the functional correctness of generated programs and provide dense feedback signals to the actor.

Le, H., Wang, Y., Gotmare, A. D., Savarese, S., & Hoi, S. C. H. Coderl: Mastering code generation through pretrained models and deep reinforcement learning. Neurips 2022

Check out our Dataset for Interactive & Reinforceme Based Grounding at AAAI 2025

POSTER PRESENTATION ON 28TH FEB

Can LLMs Obfuscate Code? A Systematic Analysis of Large Language Models into Assembly Code Obfuscation

Seyedreza Mohseni, Seyedali Mohammadi, Deepa Tilwani, Yash Saxena, Gerald Ketu Ndawula, Sriram Vema, Edward Raff, Manas Gaur

AI INSTITUTE #AIISC
UNIVERSITY OF SOUTH CAROLINA

UMBC

Grounding Verification

Despite progress in generating grounded responses, post-hoc verification of generated responses is still **indispensable**

- Especially in domains like healthcare where we may want 100% grounding
- Especially when the query is complex and/or the retrieval quality is not good
- Especially if verifiable-correct citations are required for each claim

Symbolic and Functional Grounding Together

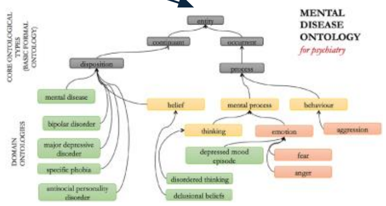
*"**Grounded generation retrieves latest clinical guidelines** and provides an evidence-based response"*

Input
I feel like the only reason I haven't already found the courage to do it is my two younger siblings. I'm sixteen, and they are twelve and ten respectively. I know that my parents would be heartbroken, but that would only be two people. My "friends" would be over it in a few weeks anyway. I feel like my brother and sister would be too emotionally damaged or traumatised if I did it, and I don't want to cause any more pain to anyone anymore.

Ground Truth Explanation: emotionally damaged, traumatised

LLAMA

I'm really sorry to hear that you're feeling this way, suicidal thoughts are serious, and it's crucial. Remember, your family care about you and want to help you through difficult times.



Domain Knowledge: PHQ9 Depression ontology

LLAMA + Domain Knowledge Output

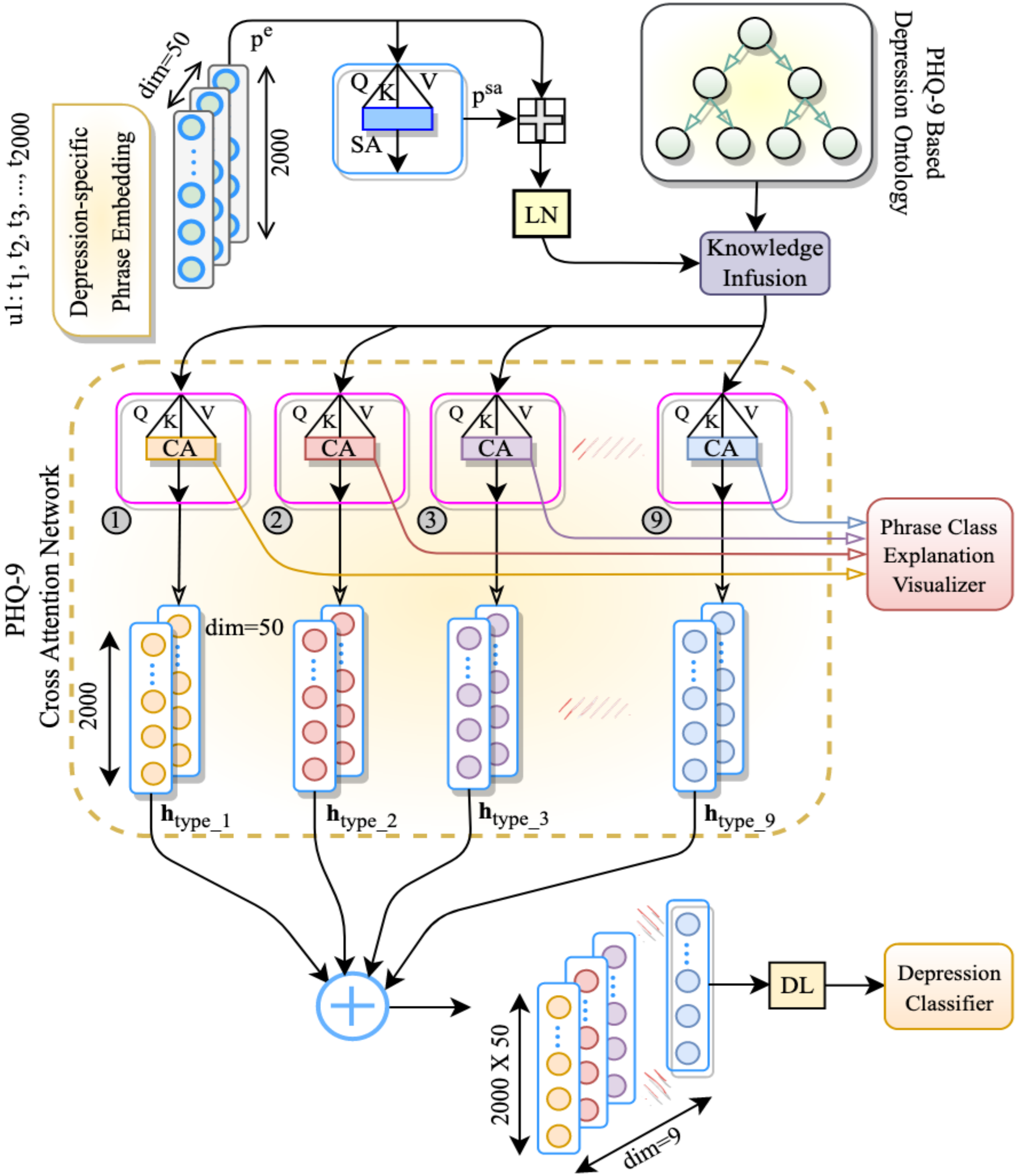
The text describes the speaker carrying a heavy burden feeling that family is *feeling irritable, emotionally disturbed and hurt*. The speaker's action *may cause more harm*.

S. Dalal, D. Tilwani, M. Gaur, S. Jain, V. L. Shalin and A. P. Sheth, "**A Cross Attention Approach to Diagnostic Explainability Using Clinical Practice Guidelines for Depression**," in *IEEE Journal of Biomedical and Health Informatics*

"Knowledge Graphs (symbolic grounding) and adapting to domain (functional grounding)"

How to do Symbolic and Functional Grounding Together ?

*“**Grounded generation retrieves latest clinical guidelines** and provides an evidence-based response”*



Original Text:

Why do i have sudden bursts of depression know the title probably doesn't make sense but stopped working for a while to peruse business idea had which failed and now i'm about go back into work force only 19 these moments where just feel lost like my family friends as is what dedicated life past 6 months most that time was me sitting in room trying get it off ground floor. really nervous getting job again haven't real one entire am overthinking or will be not bad think.

Self Attention Text *(No Highlighting)* *(Don’t know Why?)*

Why do i have sudden bursts of depression know the title probably doesn't make sense but stopped working for a while to peruse business idea had which failed and now i'm about go back into work force only 19 these moments where just feel lost like my family friends as is what dedicated life past 6 months most that time was me sitting in room trying get it off ground floor. really nervous getting job again haven't real one entire am overthinking or will be not bad think.

Attention Over PHQ 1:***How often have you been bothered by little interest or pleasure in doing things?*** ***(No Highlighting)***

Why do i have sudden bursts of depression know the title probably doesn't make sense but stopped working for a while to peruse business idea had which failed and now i'm about go back into work force only 19 these moments where just feel lost like my family friends as is what dedicated life past 6 months most that time was me sitting in room trying get it off ground floor. really nervous getting job again haven't real one entire am overthinking or will be not bad think.

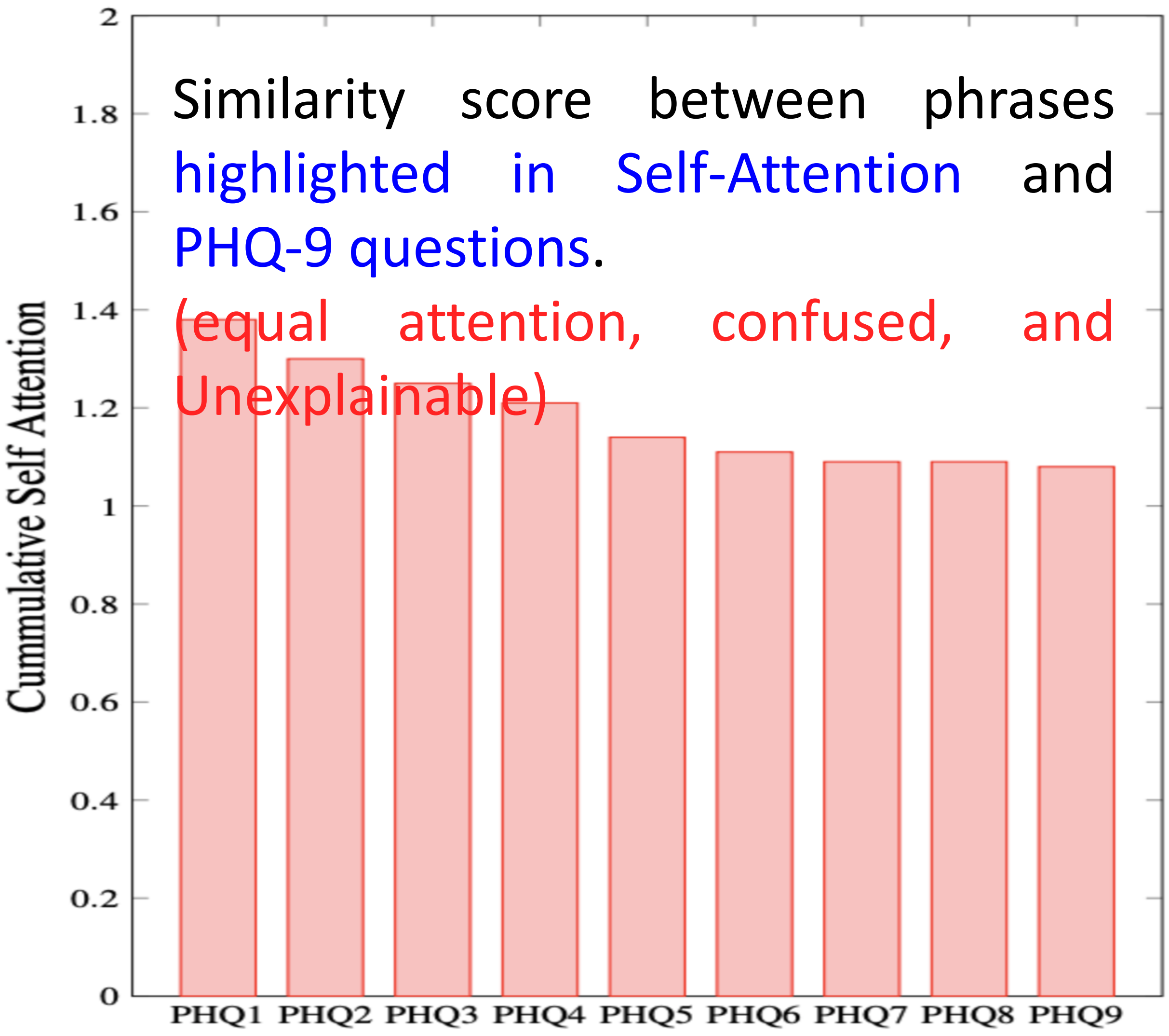
Attention Over PHQ 2 : ***How often are you bothered by feeling down, depressed, or hopeless?***

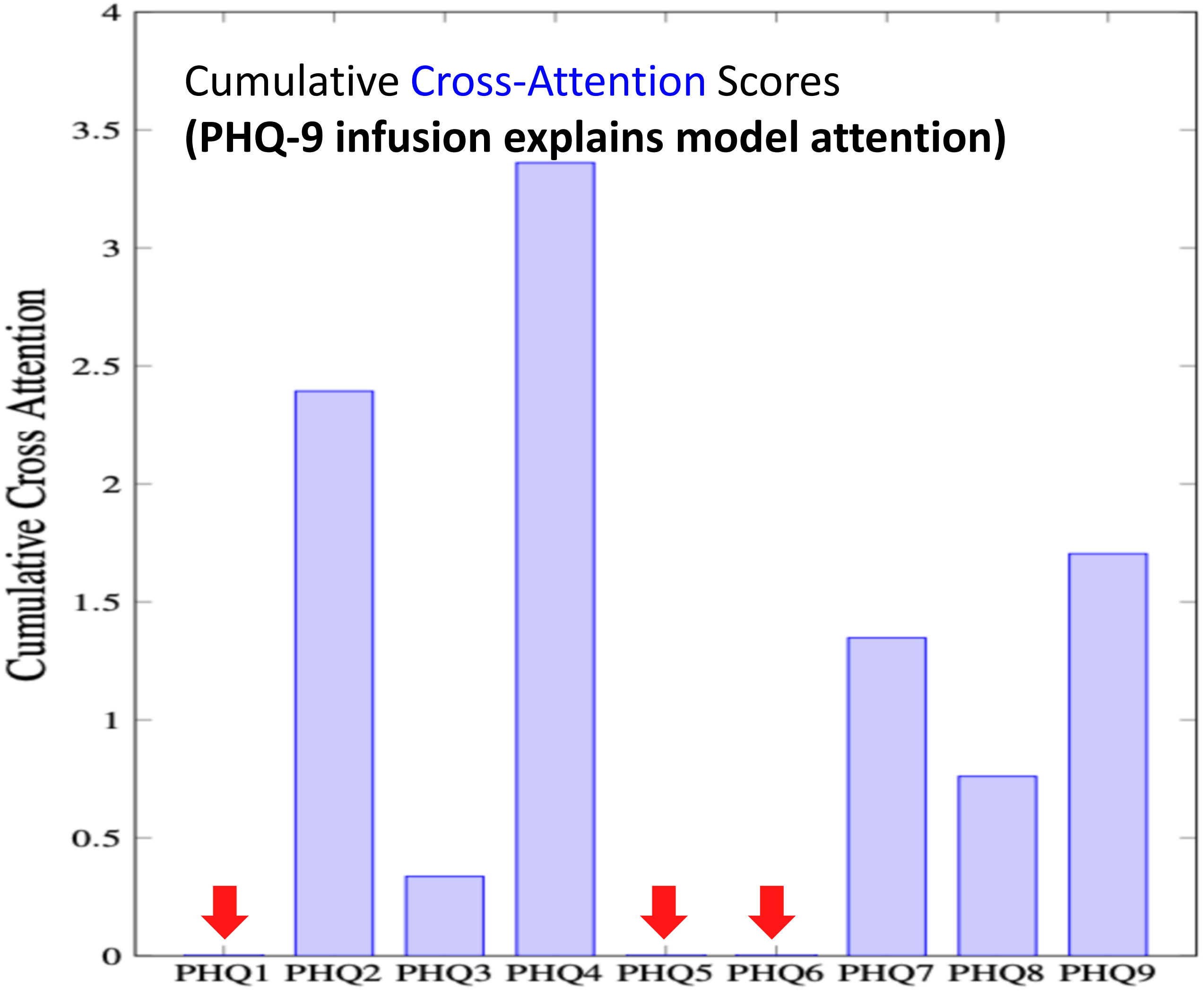
Why do i have sudden bursts of depression know the title probably doesn't make sense but stopped working for a while to peruse business idea had which failed and now i'm about go back into work force only 19 these moments where just feel lost like my family friends as is what dedicated life past 6 months most that time was me sitting in room trying get it off ground floor. really nervous getting job again haven't real one entire am overthinking or will be not bad think.

Attention Over PHQ 9: ***How often have you been bothered by thoughts that you would be better off dead or of hurting yourself in some way ?***

Why do i have sudden bursts of depression know the title probably doesn't make sense but stopped working for a while to peruse business idea had which failed and now i'm about go back into work force only 19 these moments where just feel lost like my family friends as is what dedicated life past 6 months most that time was me sitting in room trying get it off ground floor. really nervous getting job again haven't real one entire am overthinking or will be not bad think.

PHQ-1, PHQ-5, and PHQ-6 are unanswered questions. These are the relevant questions to be asked.





Check Grounding API [Google Cloud]

Check Grounding determines how grounded a given response is in a given set of facts (context)

Returns:

- Grounding scores (a support score, and a contradiction score)
- Citations
- Anti-Citations

Based on custom NLI model
Generally available at: [https://cloud.google.com/generative-\ai-app-builder/docs/check-grounding](https://cloud.google.com/generative-ai-app-builder/docs/check-grounding)

Answer candidate	Check grounding response
`Here is what I found. Titanic was directed by James Cameron.`	Support score: `0.99` Cited chunks: 0. `[From FACT 0]....Titanic is a 1997 American epic romantic disaster movie. It was directed, written, and co-produced by James Cameron. It stars Kate Winslet and Leonardo DiCaprio. The movie was released on December 19, 1997. It received positive critical reviews. The movie won 11 Academy Awards, and was nominated for fourteen total Academy Awards.....`  Claims and citations:  • 0. `Here is what I found.` Grounding check required: `false`  • 1. `Titanic was directed by James Cameron.` Citations: `[0]` Grounding check required: `true`

Open Questions..

- What mechanisms can be implemented for LLMs to **flag uncertain or unverifiable information** in their responses?
- How should LLMs handle **conflicting information** when processing text from multiple sources?
- Can LLMs dynamically **cross-reference their outputs** with primary sources or citations before finalizing a response?

Handoff

TH10: Neurosymbolic AI for EGI:

Explainable and Grounded Generations

Feb 25th 25

Ali Mohammadi, M294@umbc.edu
Ph.D. student at UMBC

University of Maryland, Baltimore County (UMBC), Knowledge Infused AI and Inference (KAI2) Lab

The 39th Annual AAAI Conference on Artificial Intelligence

FEBRUARY 25 – MARCH 4, 2025 | PHILADELPHIA, PENNSYLVANIA, USA

Key Focus Areas

Wellness Dimension

Large Language Models (LLMs)

Explainabili
ty

External Knowledge

Wellness Dimension Datasets

1. **MultiWD dataset**
2. **WellXplain dataset**

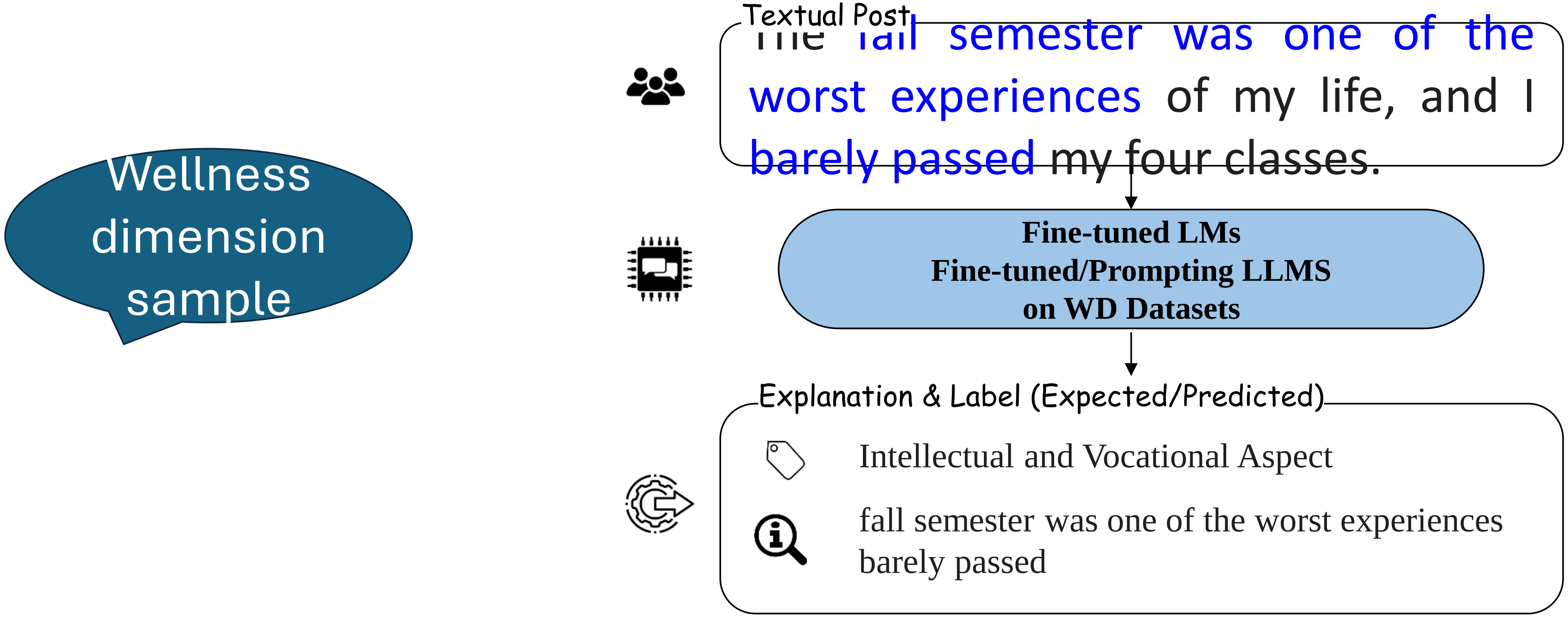
MAYO CLINIC

Substance Abuse and Mental Health Services Administration
SAMHSA
www.samhsa.gov

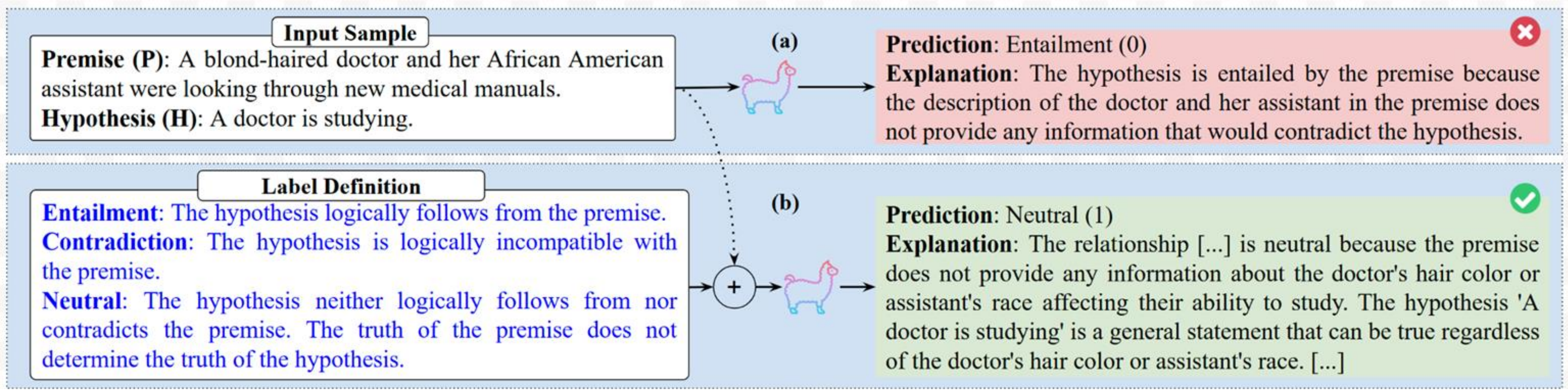
6 wellness dimensions:

- Physical
- Intellectual
- Vocational
- Social
- Spiritual
- Emotional

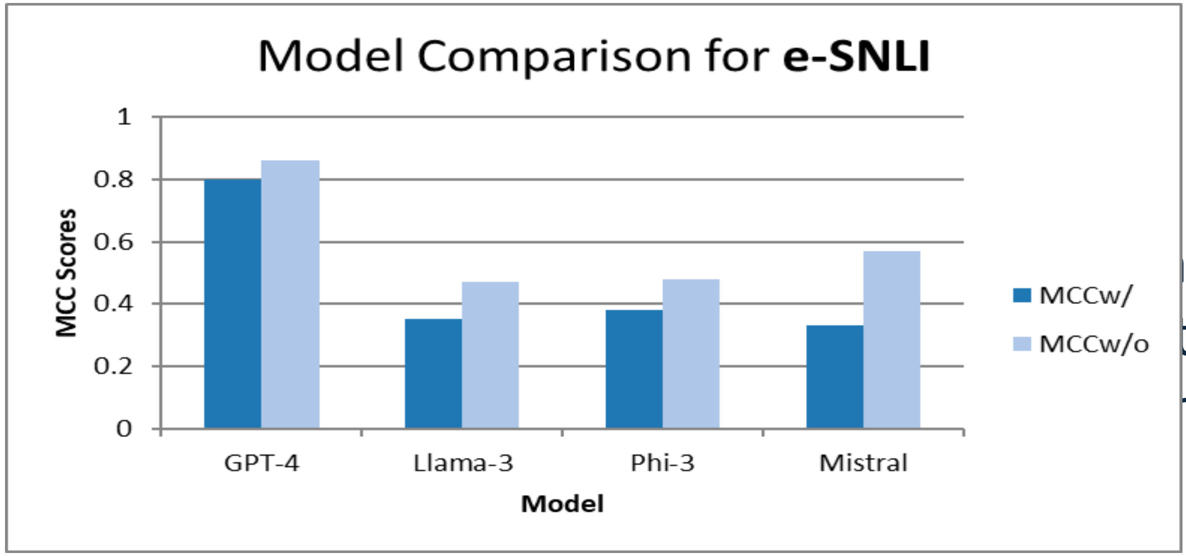
Explanation and Prediction

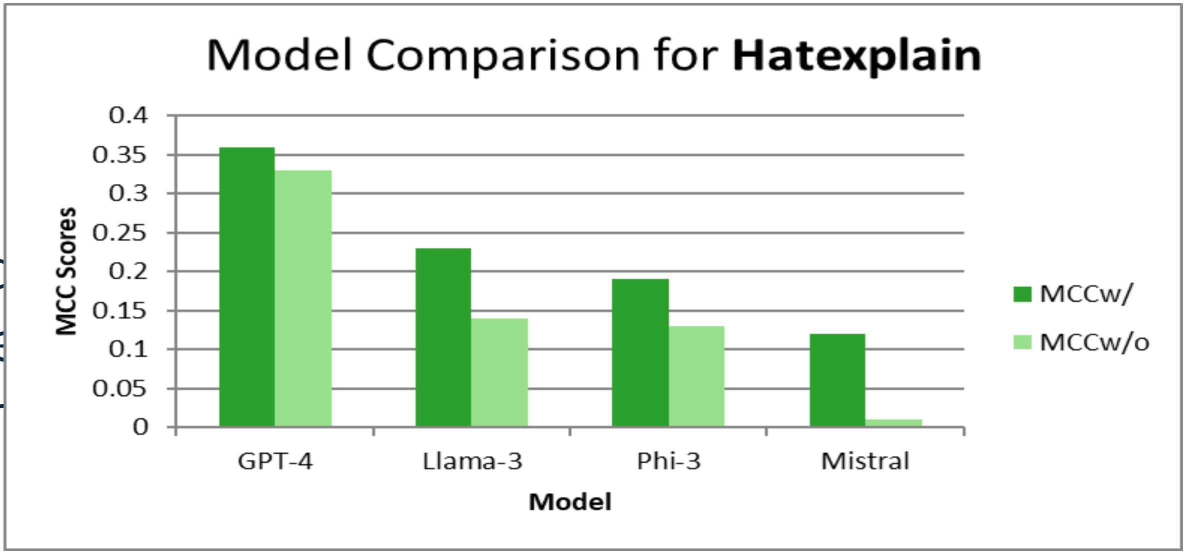


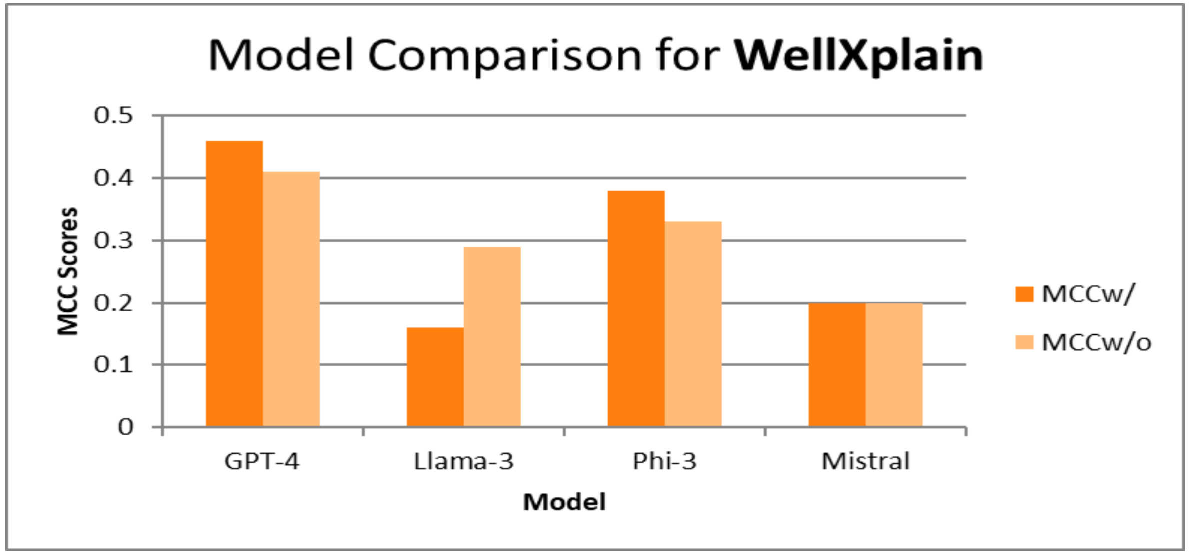
External Knowledge

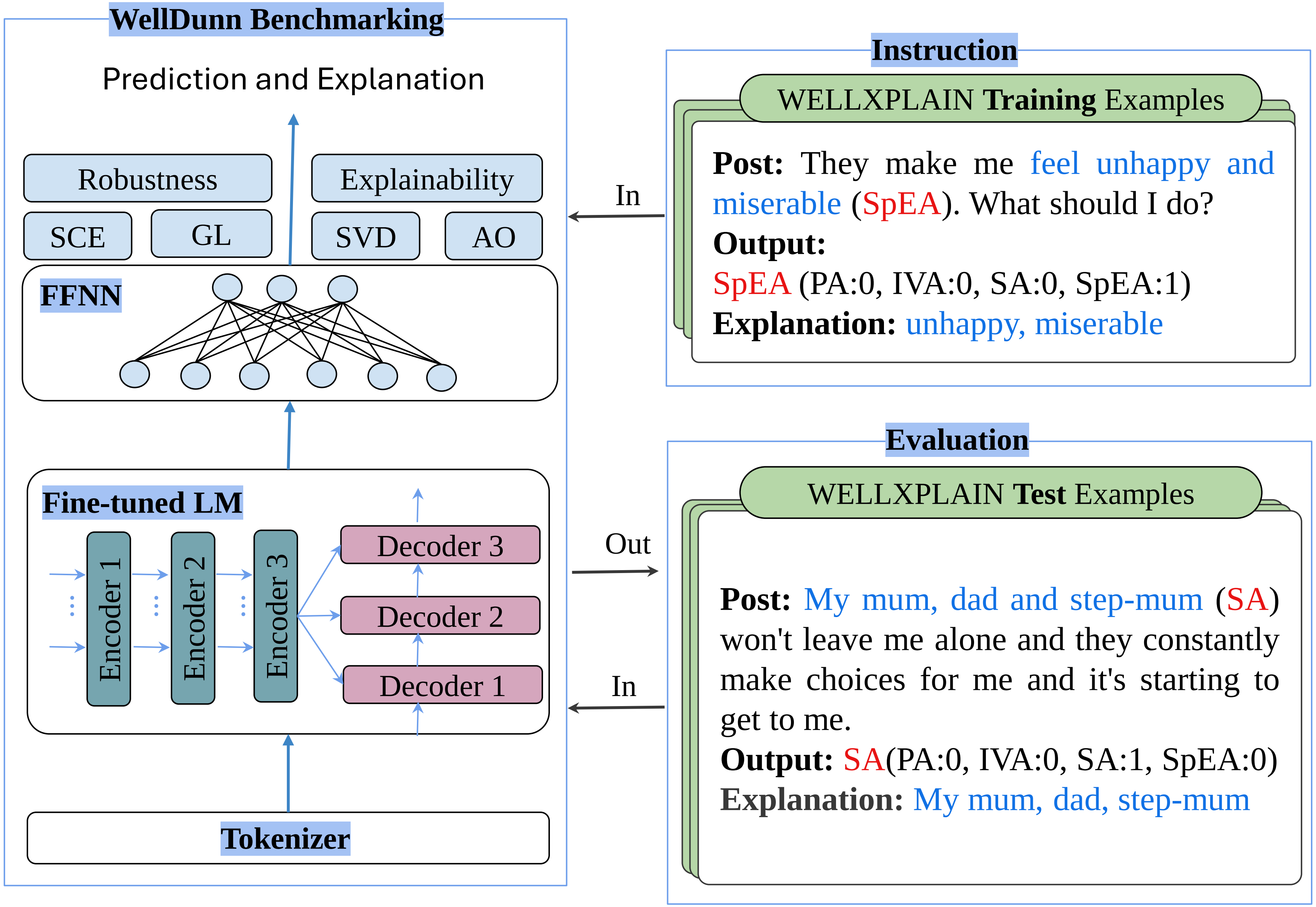


External Knowledge









Mohammadi, S., Raff, E., Malekar, J., Palit, V., Ferraro, F., & Gaur, M. (2024). WellDunn: On the Robustness and Explainability of Language Models and Large Language Models in Identifying Wellness Dimensions. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 364–388, Miami, Florida, US. ACL.

- **Robustness** Assessment
  - Sigmoid Cross-Entropy (**SCE**):
  - Gambler's Loss (**GL**):

$$SCE = -\frac{1}{n}\sum_{i=1}^{n}(y_i.\log(\sigma(x_i))+(1-y_i).\log(1-\sigma(x_i)))$$

$$GL = -\frac{1}{n}\sum_{i=1}^{n}(y_i.\log(y_i+g)); \text{where g is an abstention function in range(0,1)}$$

- **Explainability** Assessment
  - SVD (Singular Value Decomposition) ranking: measures the focus of a model's attention by analyzing its attention matrix. Lower ranks suggest that the model focuses on fewer, more relevant parts of the input text, aligning closely with concise explanations.
  - Attention-Overlap (AO) Score: The Number of Common (Purple) Words between grounded and generated explanation divided by Number of Ground-Truth Words.

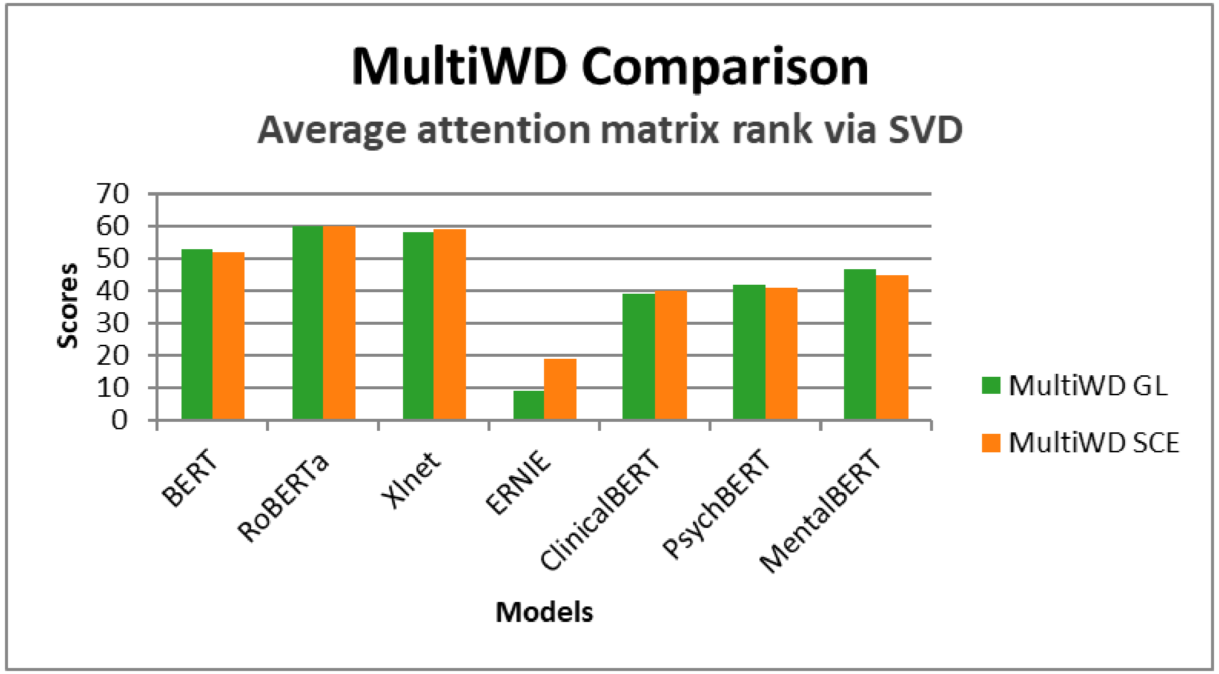
[Ground Truth Explanation]

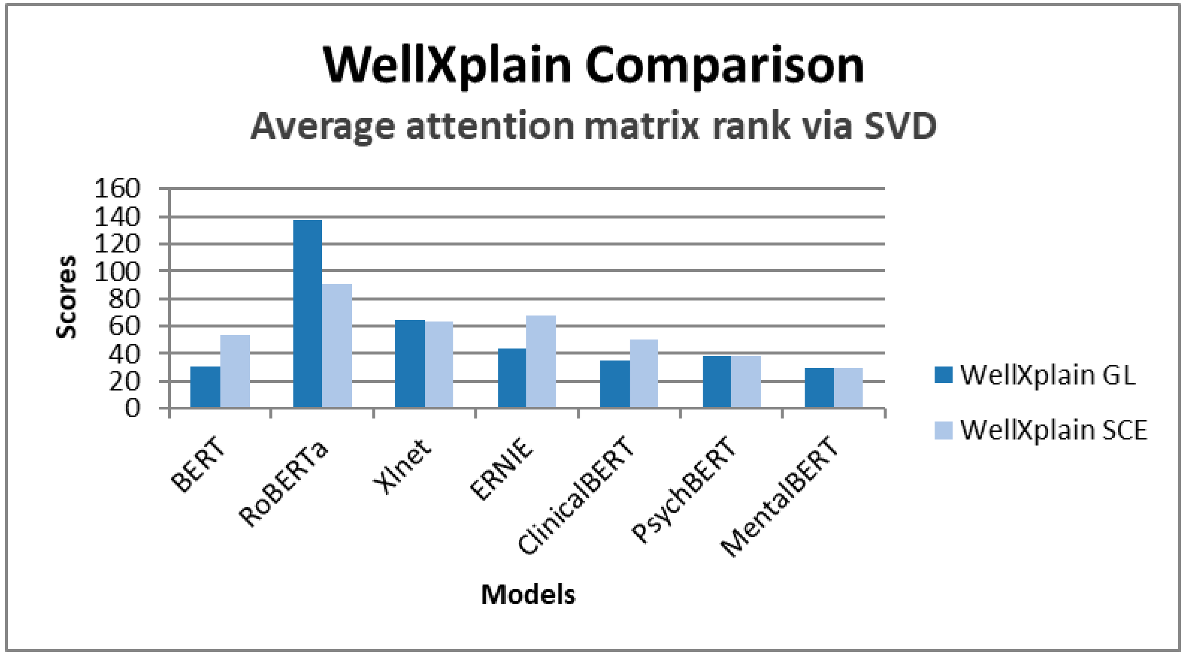
The fall semester was one of the worst experiences of my life, and I barely passed my four classes.

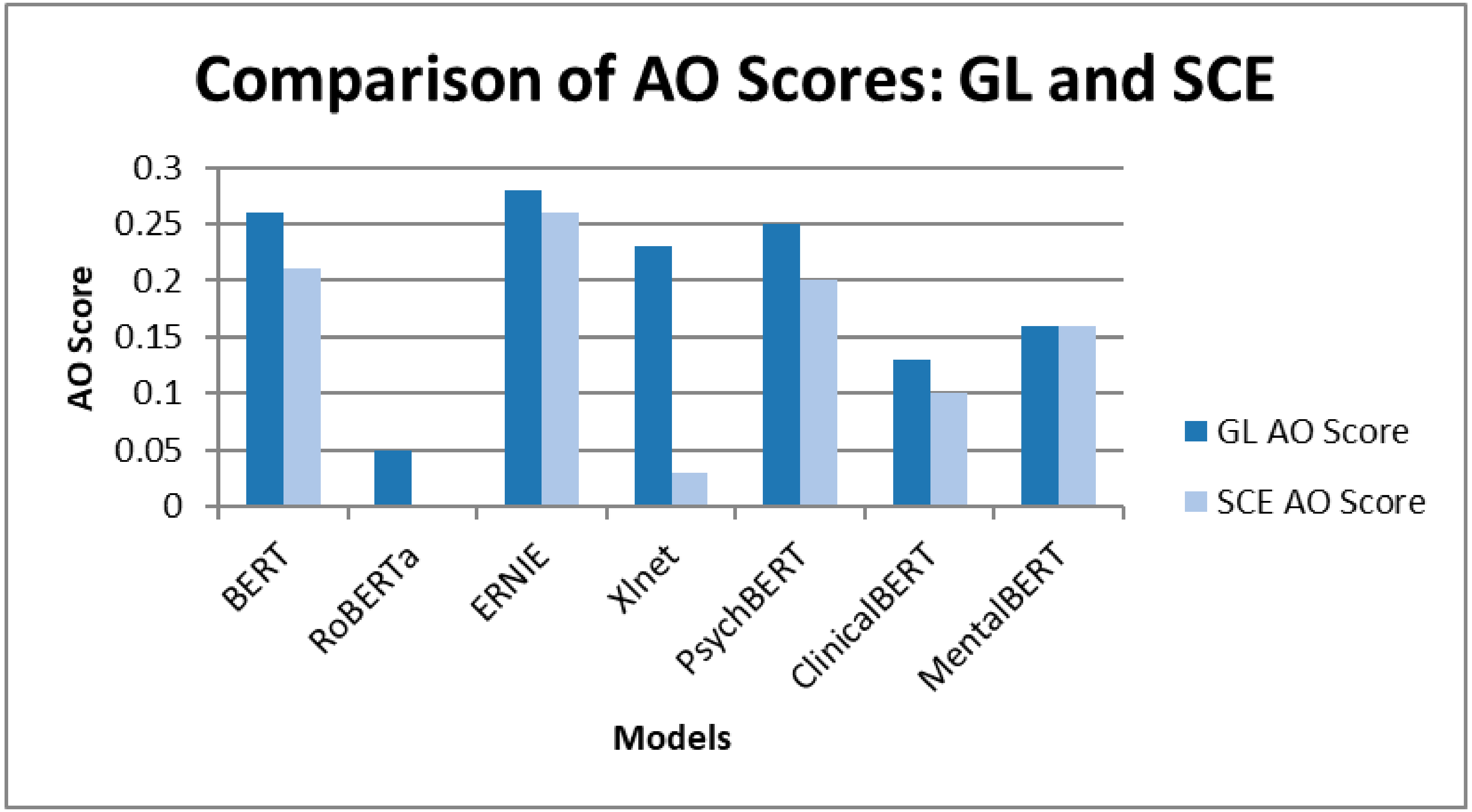
The fall semester was one of the worst experiences of my life, and I barely passed my four classes.

[Generated Explanation]

Mohammadi, S., Raff, E., Malekar, J., Palit, V., Ferraro, F., & Gaur, M. (2024). WellDunn: On the Robustness and Explainability of Language Models and Large Language Models in Identifying Wellness Dimensions. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 364–388, Miami, Florida, US. ACL.


MultiWD Comparison
Average attention matrix rank via SVD
Scores
70
60
50
40
30
20
10
0
BERT
RoBERTa
Xlnet
ERNIE
ClinicalBERT
PsychBERT
MentalBERT
Models
MultiWD GL
MultiWD SCE


WellXplain Comparison
Average attention matrix rank via SVD
Scores
160
140
120
100
80
60
40
20
0
BERT
RoBERTa
Xlnet
ERNIE
ClinicalBERT
PsychBERT
MentalBERT
Models
WellXplain GL
WellXplain SCE


Comparison of AO Scores: GL and SCE
AO Score
0.3
0.25
0.2
0.15
0.1
0.05
0
BERT
RoBERTa
ERNIE
XInet
PsychBERT
ClinicalBERT
MentalBERT
Models
GL AO Score
SCE AO Score

SCE vs GL attention (Post 1)

With SCE Loss:

I don't cry anymore. want to be around anyone do anything Work keeps me getting up everyday Without it would probably stare at my ceiling until passed back out again so tired know if there is a question in this, There just isn't else tell.

With GL:

I don't cry anymore. want to be around anyone, do anything. Work keeps me getting up everyday. Without it would probably stare at my ceiling until passed back out again m so tired know if there is a question in this, There just isn else tell.

Future Directions

Developing a Transparent Classifier Rooted in Clinical Understanding – Addressing the disparities between prediction accuracy and attention.

Improving Attention Alignment with Ground Truth – Enhancing attention explanations to better reflect actual outcomes.

Exploring Different Prompting and Retrieval-Augmented Generation (RAG) Strategies – Testing alternative methods to improve LLM performance.

Developing a Suitable Dataset for Mental Health Applications – Curating knowledge and constructing a well-suited dataset for retrieval-augmented methods.

aclanthology.org/2024.blackboxnlp-1.23/

ACL Anthology News FAQ Corrections Submissions Github

Search...

WellDunn: On the Robustness and Explainability of Language Models and Large Language Models in Identifying Wellness Dimensions

Seyedali Mohammadi, Edward Raff, Jinendra Malekar, Vedant Palit, Francis Ferraro, Manas Gaur

PDF

Cite

Search

Fix data

Abstract

Language Models (LMs) are being proposed for mental health applications where the heightened risk of adverse outcomes means predictive performance may not be a sufficient litmus test of a model's utility in clinical practice. A model that can be trusted for practice should have a correspondence between explanation and clinical determination, yet no prior research has examined the attention fidelity of these models and their effect on ground truth explanations. We introduce an evaluation design that focuses on the robustness and explainability of LMs in identifying Wellness Dimensions (WDs). We focus on two existing mental health and well-being datasets: (a) Multi-label Classification-based MultiWD, and (b) WellXplain for evaluating attention mechanism veracity against expert-labeled explanations. The labels are based on Halbert Dunn's theory of wellness, which gives grounding to our evaluation. We reveal four surprising results about LMs/LLMs: (1) Despite their human-like capabilities, GPT-3.5/4 lag behind RoBERTa, and MedAlpaca, a fine-tuned LLM on WellXplain fails to deliver any remarkable improvements in performance or explanations. (2) Re-examining LMs' predictions based on a confidence-oriented loss function reveals a significant performance drop. (3) Across all LMs/LLMs, the alignment between attention and explanations remains low, with LLMs scoring a dismal 0.0. (4) Most mental health-specific LMs/LLMs overlook domain-specific knowledge and undervalue explanations, causing these discrepancies. This study highlights the need for further research into their consistency and explanations in mental health and well-being.

Anthology ID: 2024.blackboxnlp-1.23

Volume: Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP

Month: November

Year: 2024

Reference

1. Seyedali Mohammadi, Edward Raff, Jinendra Malekar, Vedant Palit, Francis Ferraro, and Manas Gaur. 2024. WellDunn: On the Robustness and Explainability of Language Models and Large Language Models in Identifying Wellness Dimensions. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 364–388, Miami, Florida, US. Association for Computational Linguistics.
2. Liu, Ziyin, et al. "Deep gamblers: Learning to abstain with portfolio theory." *Advances in Neural Information Processing Systems* 32 (2019).
3. Alsentzer, Emily, et al. "Publicly available clinical BERT embeddings." *arXiv preprint arXiv:1904.03323* (2019).
4. WELLDUNN: An Annotated Dataset to identify affected Wellness Dimensions in Reddit Posts, Under review.
5. DUNN HL. High-level wellness for man and society. Am J Public Health Nations Health. 1959 Jun;49(6):786-92. doi: 10.2105/ajph.49.6.786. PMID: 13661471; PMCID: PMC1372807.
6. Merikangas KR, He JP, Burstein M, Swanson SA, Avenevoli S, Cui L, Benjet C, Georgiades K, Swendsen J. Lifetime prevalence of mental disorders in U.S. adolescents: results from the National Comorbidity Survey Replication--Adolescent Supplement (NCS-A). J Am Acad Child Adolesc Psychiatry. 2010 Oct;49(10):980-9. doi: 10.1016/j.jaac.2010.05.017. Epub 2010 Jul 31. PMID: 20855043; PMCID: PMC2946114.
7. Garg, M. Mental Health Analysis in Social Media Posts: A Survey. *Arch Computat Methods Eng* 30, 1819–1842 (2023). https://doi.org/10.1007/s11831-022-09863-z
8. Garg, Muskan. "WellXplain: Wellness concept extraction and classification in Reddit posts for mental health analysis." *Knowledge-Based Systems* 284 (2024): 111228.
9. Sathvik, M. S. V. P. J., and Muskan Garg. "Multiwd: Multiple wellness dimensions in social media posts." *Authorea Preprints* (2023)

openCHA:
Building Explainable and Personalized Conversational Agent

Healthcare chatbots or Conversational Health Agents

Chatbots have the potential to play a crucial role in healthcare: assisting patients and healthcare providers:

- Clinical decision support
- Patient monitoring and follow-up
- Chronic health management
- Patient’s self-awareness

Bedi, Suhana, et al. "A Systematic Review of Testing and Evaluation of Healthcare Applications of Large Language Models (LLMs)." medRxiv (2024): 2024-04.

Why are healthcare chatbots not widely used?

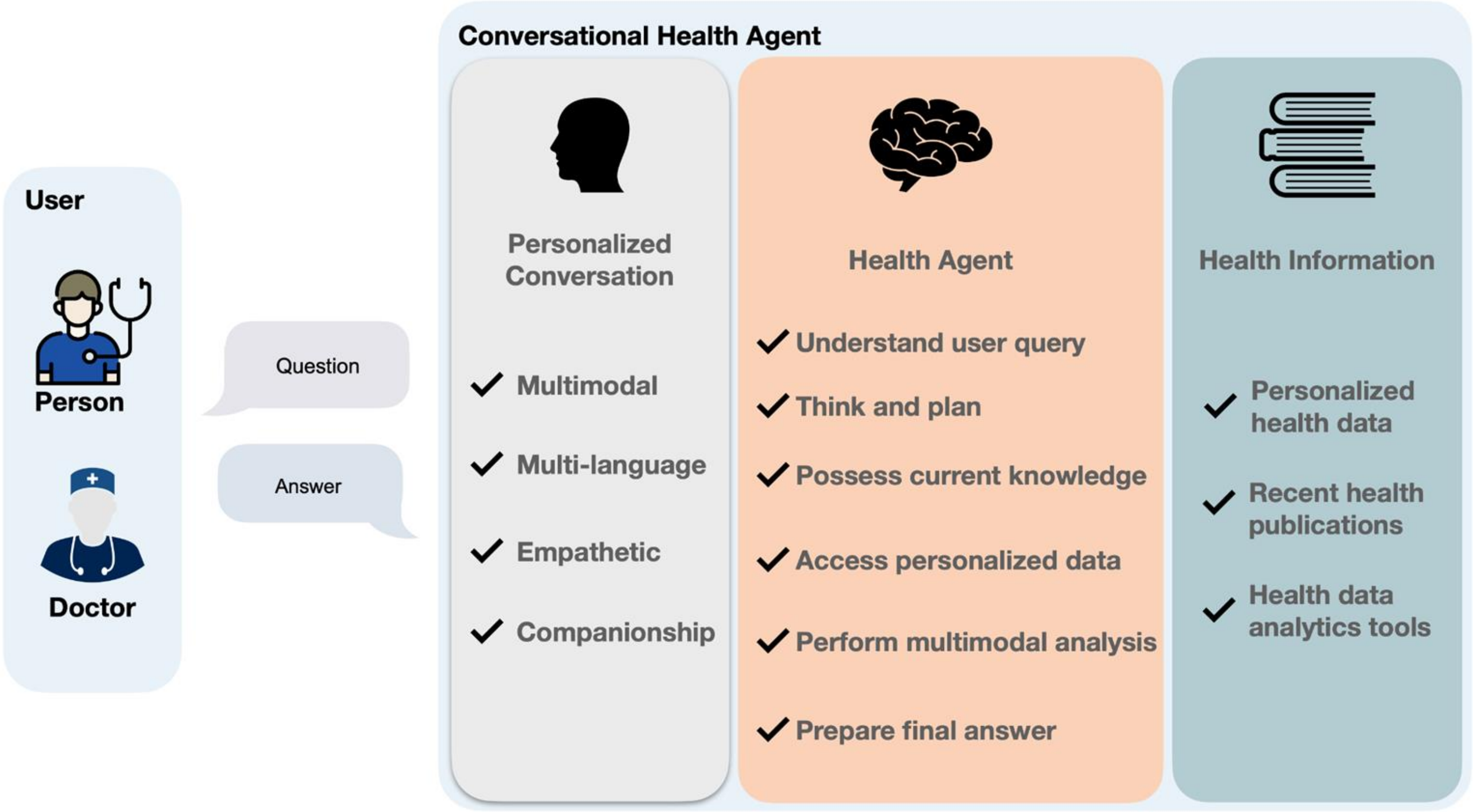
Existing chatbots are not able to provide:

1. Trustworthiness
2. Personalization (no access to user's data)
3. Data analysis
   - No access to established multimodal data analysis tools
4. Access to up-to-date information
   - Ignoring well-established healthcare research
5. Explainability
6. ...

Abbasian, M., Khatibi, E., Azimi, I., Oniani, D., Shakeri Hossein Abad, Z., Thieme, A., ... & Rahmani, A. M. (2024). Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. npj Digital Medicine, 7(1), 82.

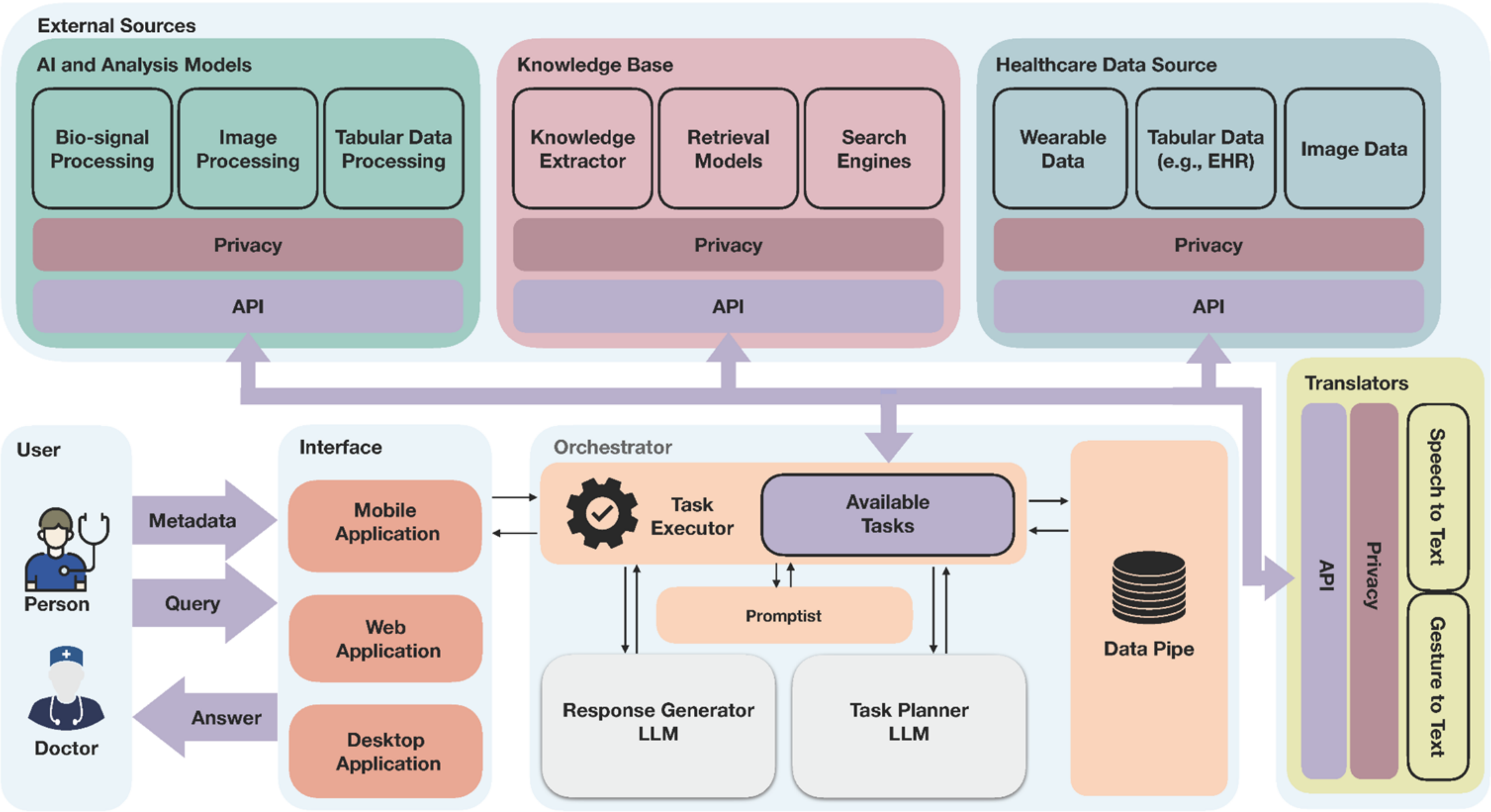
openCHA (Conversational Health Agents)

A holistic LLM-powered framework to integrate **health data, knowledge,** and **analytical tools** into healthcare **chatbots.**



- Abbasian, M., Azimi, I., Rahmani, A.M. and Jain, R., 2023. Conversational Health Agents: A Personalized LLM-Powered Agent Framework. arXiv preprint arXiv:2310.02374.
- GitHub repo: https://github.com/Institute4FutureHealth/CHA

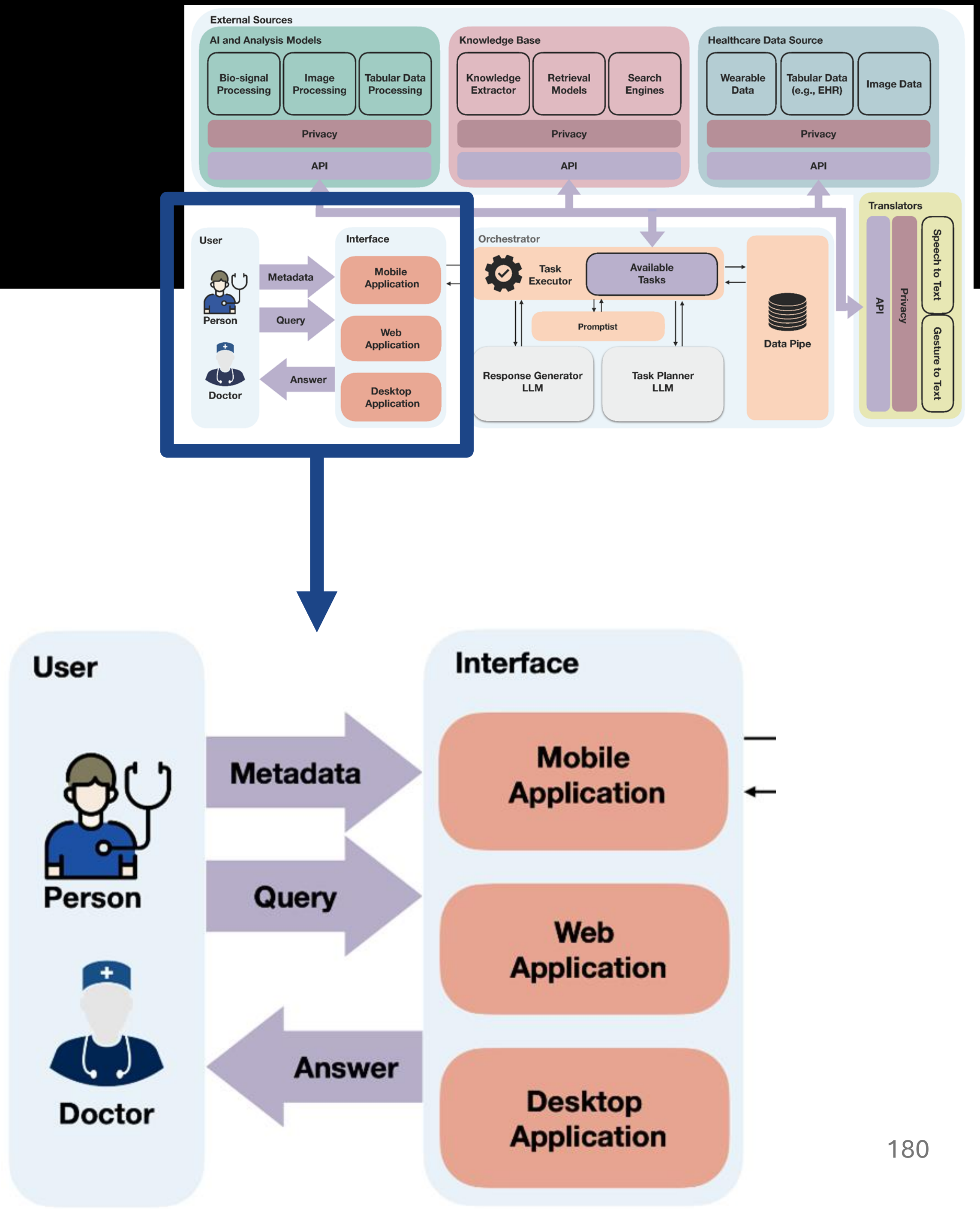
openCHA framework



Interface

Acts as a bridge between the users and agents

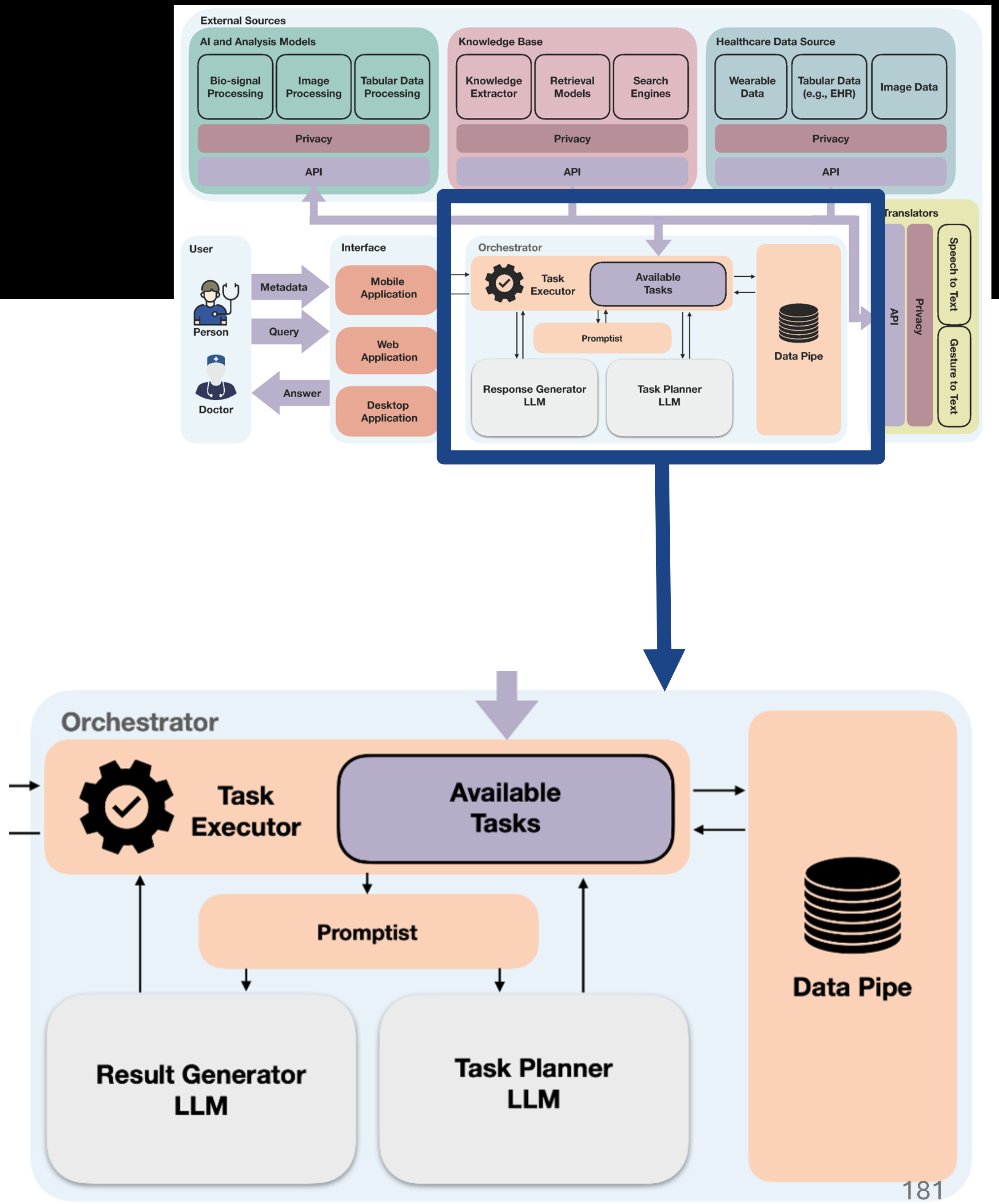
- Users' queries
- Metadata



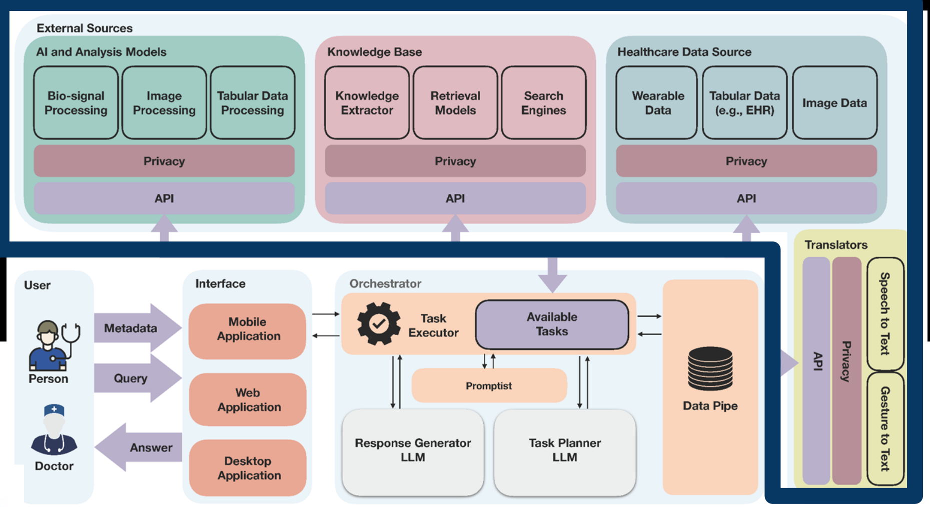
Orchestrator

Responsible for problem solving, decision making, and response generation

- Input data are aggregated, transformed into structured data, and then analyzed to plan and execute actions
- Interacts with external sources to acquire the required information, perform data integration and analysis, and extract insights, among other functions.
- Converts the info into an understandable format and inferring the appropriate response.



External sources



Obtain essential information from the broader world

- Datasets
- Knowledge bases
- Analytical tools
- Translators

Demo:
Nutrition causal effects

Tasks involved:

- Get data
- Causal graph (personal info)
- Food’s nutritional content (general info)

Z. Yang, E. Khatibi, N. Nagesh,, M. Abbasian, I. Azimi, R. Jain, and A. Rahmani, “ChatDiet: Empowering Personalized Nutrition-Oriented Food Recommender Chatbots through an LLM-Augmented Framework,” Elsevier Smart Health, IEEE/ACM CHASE, 2024

Patient health record reporting (1)

Tasks involved:

- Get data
- Statistical analysis
- Internet search
- Extract text

Is REM sleep of Patient 5 sufficient during August 2020?

Based on the data collected from Patient 5 in August 2020, the REM sleep duration was 110.63 minutes. In order to determine if this is sufficient, we need to consider the individual's specific sleep needs and overall sleep patterns. It's important to note that the optimal amount of REM sleep can vary from person to person, but generally, adults need about 90-120 minutes of REM sleep each night.

In this case, the REM sleep duration of 110.63 minutes falls within the typical range for adults. However, it's important to assess the overall sleep quality and patterns as well. Factors such as sleep efficiency and total sleep time also play a role in determining if the REM sleep is sufficient for the individual's needs.

For a more holistic view of the patient's sleep health, it would be beneficial to consider other sleep metrics such as total sleep time, sleep efficiency, and sleep onset latency, in addition to the REM sleep duration. This can provide a more comprehensive understanding of the patient's sleep patterns and overall sleep health.

For further analysis or to better understand the significance of the REM sleep duration in relation to the patient's overall sleep quality, it may be helpful to consult a healthcare professional or sleep specialist for personalized guidance and recommendations.

Dataset: S. Labbaf, et al. "Physiological and Emotional Assessment of College Students Using Wearable and Mobile Devices During the 2020 Covid-19 Lockdown: An Intensive, Longitudinal Dataset." Longitudinal Dataset (2023).

Patient health record reporting (2)

Tasks involved:

- Get data
- Statistical analysis
- SerpAPI
- Extract Text

Give me reference on the REM sleep recommendation

According to the WHOOP, Rapid Eye Movement (REM) sleep is an essential stage of sleep that has important physiological and cognitive functions. It is known as the "mentally restorative" stage of sleep when the brain converts short-term memories into long-term ones. The average adult should aim to spend around 20-25% of their total sleep time in the REM stage, with around 90 minutes of REM sleep being recommended for individuals who get 7-8 hours of sleep. This means that the REM sleep duration of 110.63 minutes for Patient 5 in August 2020 falls within the typical range for adults.

However, REM sleep needs can vary from person to person, and other factors such as overall sleep patterns, sleep efficiency, and total sleep time also play a role in determining if the REM sleep is sufficient for an individual's needs. It's important to consider a more comprehensive view of the patient's sleep health by assessing other sleep metrics such as total sleep time, sleep efficiency, and sleep onset latency.

Ultimately, for a more in-depth analysis of the significance of the REM sleep duration in relation to the patient's overall sleep quality, it may be helpful to consult a healthcare professional or sleep specialist for personalized guidance and recommendations.

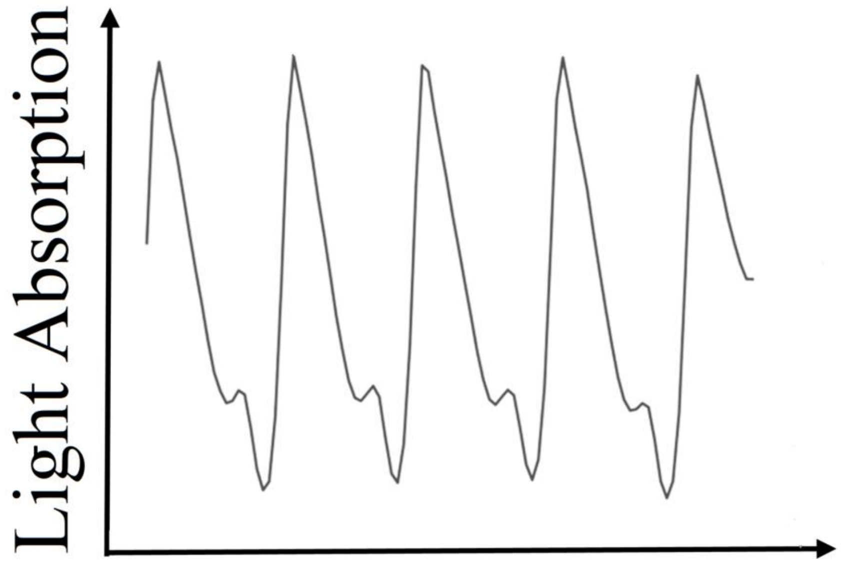
If you would like to learn more, you can visit the following link for additional information on REM sleep: WHOOP - What is REM Sleep, Why it's Important, How Much You Need & How to Increase It.

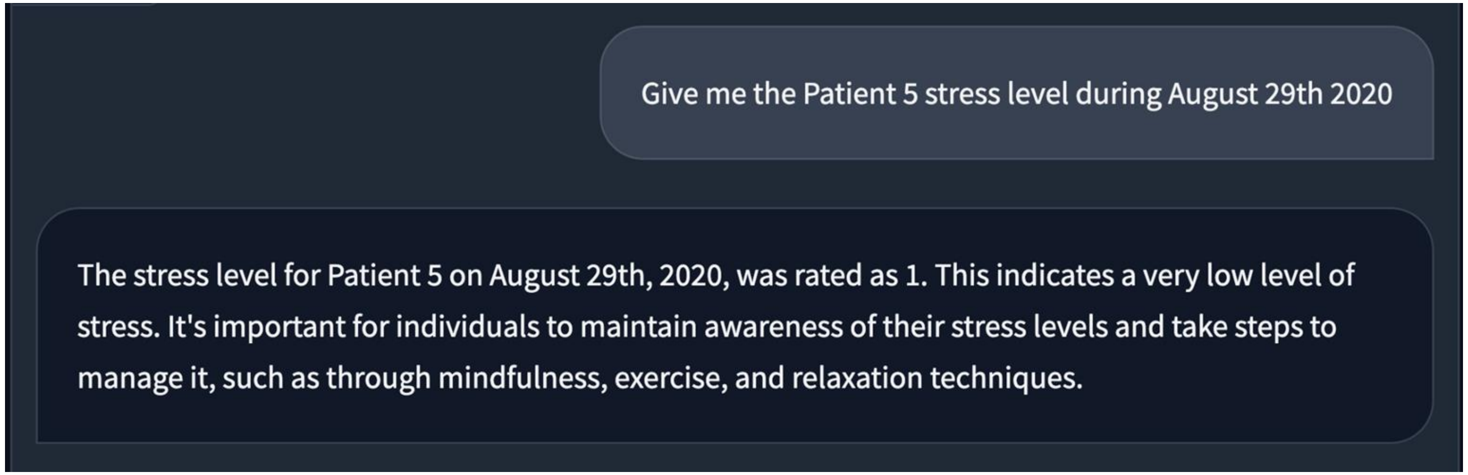
Dataset: S. Labbaf, et al. "Physiological and Emotional Assessment of College Students Using Wearable and Mobile Devices During the 2020 Covid-19 Lockdown: An Intensive, Longitudinal Dataset." Longitudinal Dataset (2023).

Objective stress level estimation

Tasks involved:

- Get data
- PPG analysis (HRV extraction)
- HRV analysis (stress estimation)





Dataset: S. Labbaf, et al. "Physiological and Emotional Assessment of College Students Using Wearable and Mobile Devices During the 2020 Covid-19 Lockdown: An Intensive, Longitudinal Dataset." Longitudinal Dataset (2023).

Use cases

- Yang, Zhongqi, et al. "ChatDiet: Empowering personalized **nutrition-oriented food recommender** chatbots through an LLM-augmented framework." Smart Health 32 (2024): 100465.
- Abbasian, Mahyar, et al. "Knowledge-Infused LLM-Powered Conversational Health Agent: A Case Study for Diabetes Patients." 46th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), (2024).
- Park, Jung In, et al. "Building Trust in Mental Health Chatbots: Safety Metrics and **LLM-Based Evaluation** Tools." arXiv preprint arXiv:2408.04650 (2024).
- Abbasian, Mahyar, et al. "Empathy Through **Multimodality** in Conversational Interfaces." arXiv preprint arXiv:2405.04777 (2024).

Future directions

We are looking for contribution from diverse communities: to contribute their ideas and connect their tools to CHA, leading to more precise user responses.

- Safety of the responses
    - Benchmarking
- Connecting open datasets, knowledge graphs, etc.
    - Planning and decision making
    - Retrieve information
- Chronic Health Management

More info about openCHA:

arxiv.org/abs/2310.02374

GitHub repository:

github.com/Institute4FutureHealth/CHA

User guide and quick start:

opencha.com

Should you be interested, please reach out to me at

- manas@umbc.edu
- edraffi@umbc.edu
- dtilwani@mailbox.sc.edu
- m294@umbc.edu
- azimii@uci.edu

Thank You

Questions?

Slides Available : https://nesy-egi.github.io

Thanks! Questions?

- Feedback most welcome :-)
    - manas@umbc.edu
    - edraff1@umbc.edu
    - **dtilwani@mailbox.sc.edu**
    - **m294@umbc.edu**
    - azimi.iman.1988@gmail.com

- Tutorial website: https://nesy-egi.github.io