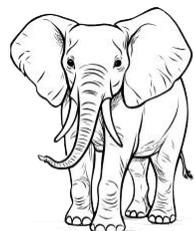




清华大学  
Tsinghua University



**ACL 2024**

*Bangkok, Thailand*

# ACL 2024 Sharing

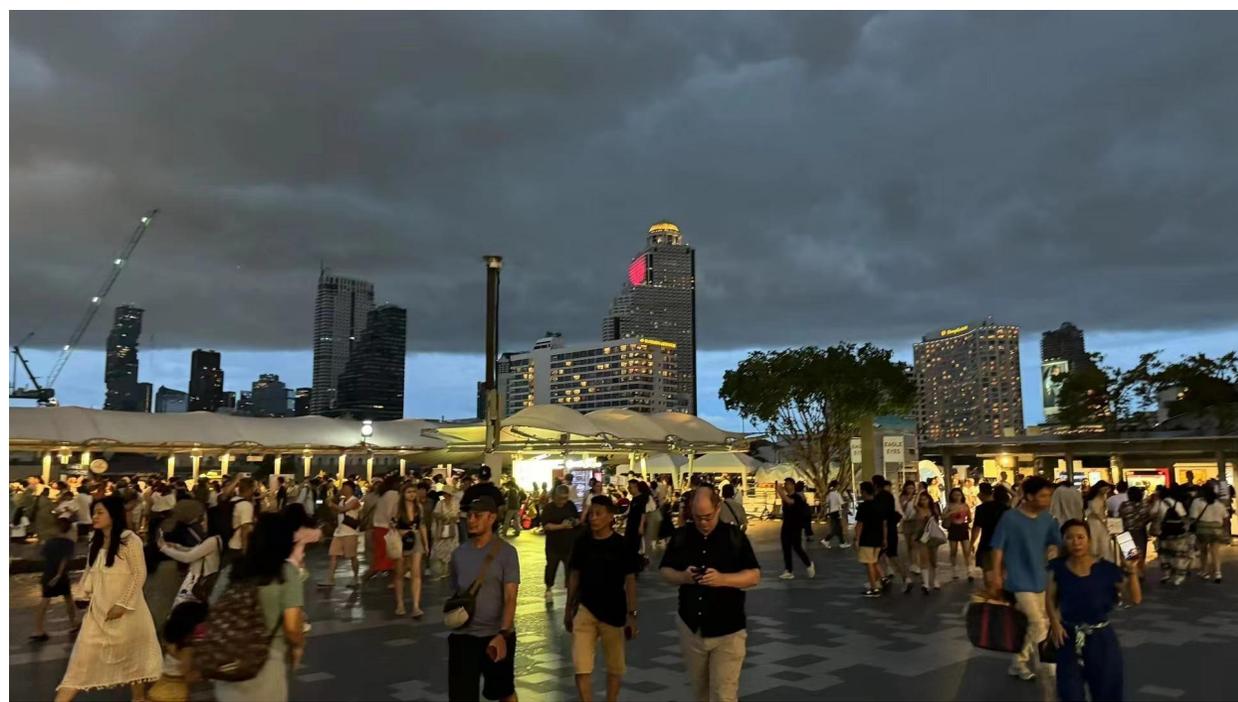
Zhu Liu

2024.09.26

# ACL 2024

- 会议时间：8月11日到16日  
(17,18 还有2天workshop)
- 会议地点：泰国曼谷的  
Centara Grand and Bangkok  
Convention Centre at  
CentralWorld





# 曼谷整体印象

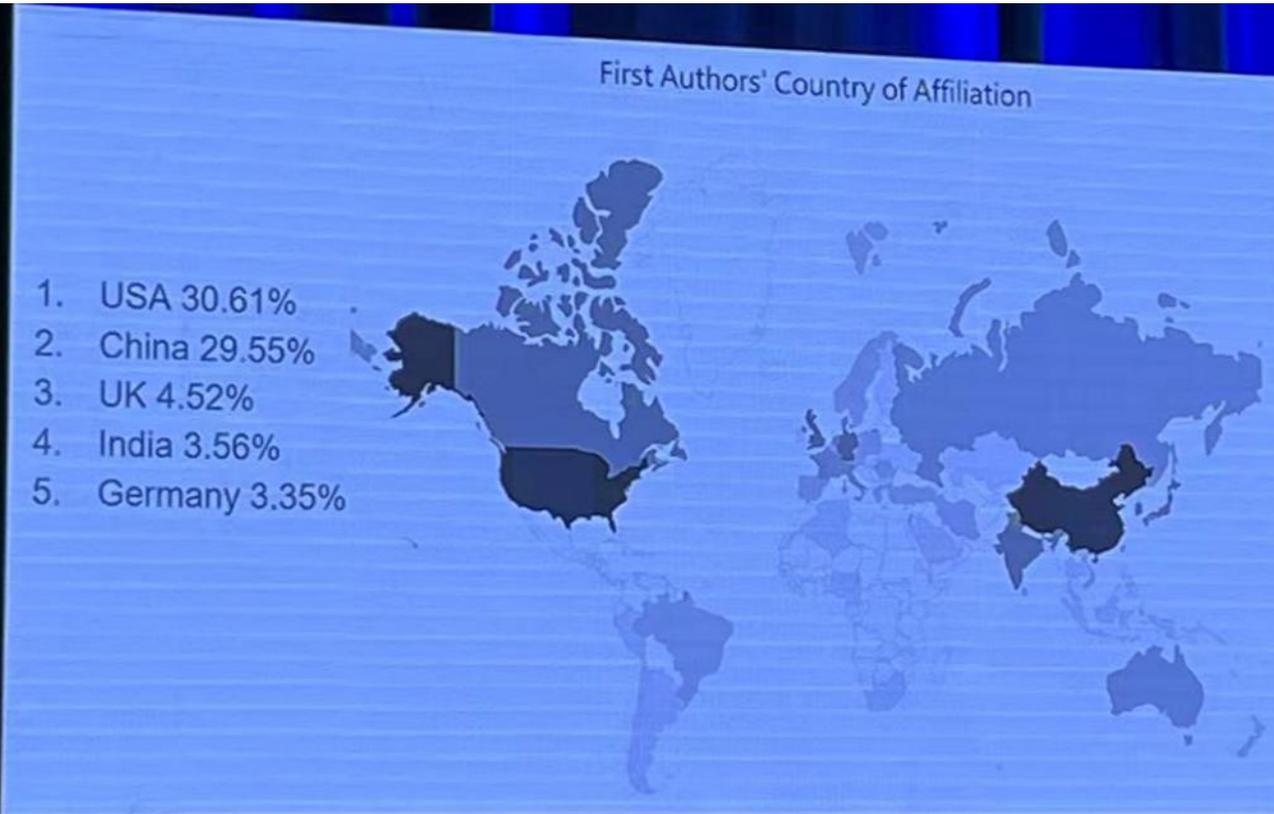
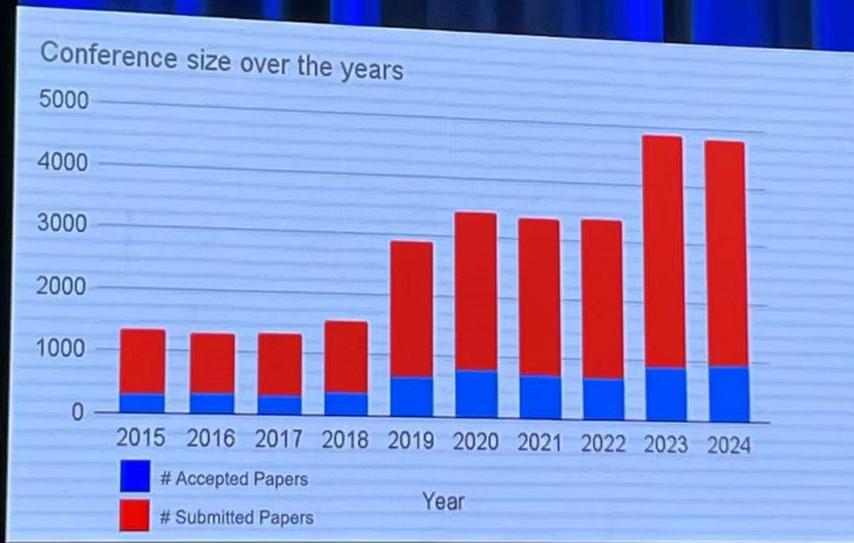
- 传统和现代的交融
  - 传统：很多佛教元素、保留王室、神秘的建筑
  - 现代：高楼大厦、便捷的购物体验...
- 天气湿热
- 饮食偏辛辣
- 一些可能吸引办会的点：国际化程度高、交通便捷、签证方便、英语普及度较高、很安全

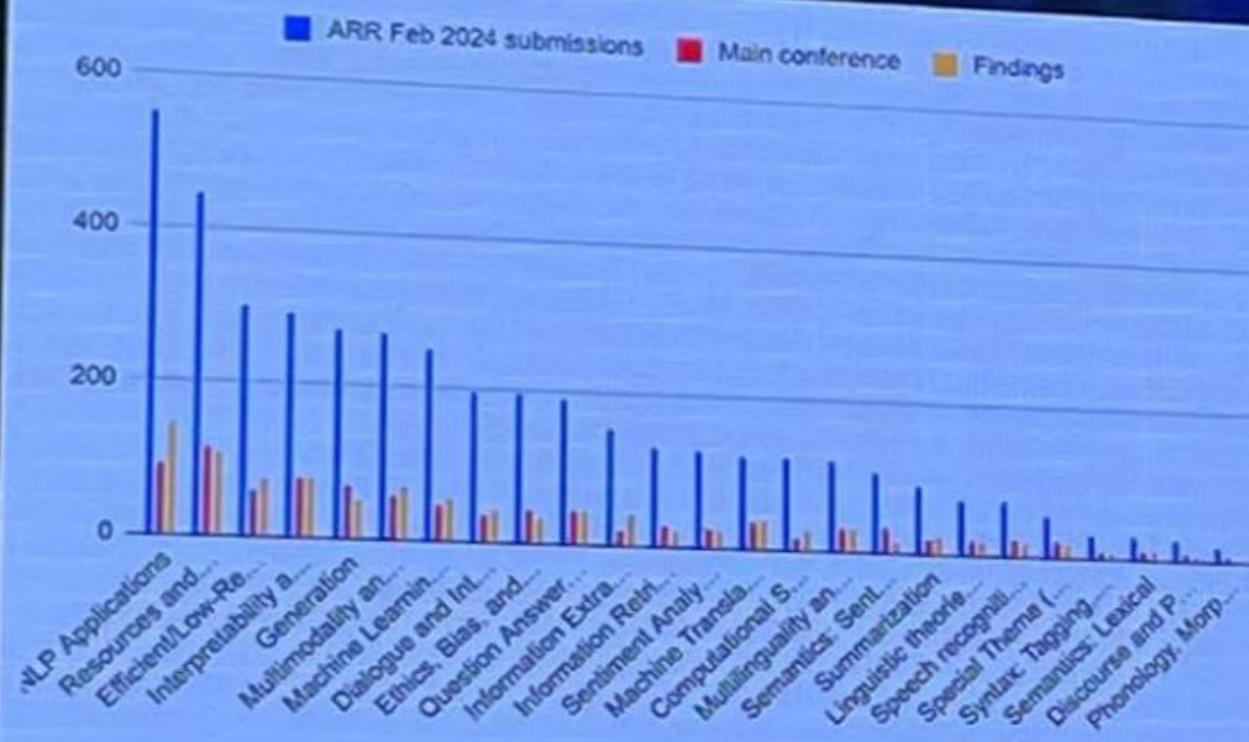
# 议程

日期	时间	活动内容
8月10日	14:00-19:00	注册
8月11日	07:30-20:30	注册、教程、欢迎招待会
8月12日	07:30-17:30	注册、主题演讲、口头报告/海报/演示
8月13日	08:30-16:30	注册、主题演讲、圆桌讨论、社交晚宴
8月14日	08:30-16:30	注册、主题演讲、奖项与闭幕
8月15-16日	08:00-16:30	注册、研讨会

# ACL 2024 介绍

- 主会
  - 约5000篇投稿, 72个SAC, 716个AC, 4208个审稿人
  - 940篇main (21.3%) , 975篇findings (22.1%) , 6 JCL, 31 TACL
  - 3个主旨演讲, 1个圆桌讨论
- 其他
  - 18个workshop, 6个教程, 38个demos和60篇SRW的论文





**Largest tracks this year:**

- NLP Applications
- Resources and Evaluation
- Efficient/Low-Resource Methods for NLP

- 2023: NLP Applications, Machine Learning, Information Extraction
- 2022: Machine Learning, Information Extraction, NLP Applications
- 2021: Machine Learning, Machine Translation & Multilinguality, Information Extraction

## What's New This Year



- **All papers** presented as posters, some papers also as orals
- All findings papers can present (as posters)
- Virtual day **after** the conference (next week: Thursday, August 22)
- Non-publicized paper award
- Theme: Open science, open data, and open models for reproducible NLP

# 主旨演讲

- Does In-Context-Learning Offer the Best Tradeoff in Accuracy, Robustness, and Efficiency for Model Adaptation?
- Can LLMs Reason and Plan?
- Are LLMs Narrowing Our Horizon? Let's Embrace Variation in NLP!

# 主旨演讲

- Does In-Context-Learning Offer the Best Tradeoff in Accuracy, Robustness, and Efficiency for Model Adaptation?
- 主讲人: Sunita Sarawagi
  - Professor, IIT Bombay
  - sequence models for text and time-series, domain adaptation, effective human intervention in learning, graphical models and structured learning



# Background

- The model adaptation problem: from one domain (training set) to another (test set)
- Challenges:
  - Accuracy (overfit vs. no change)
  - Robustness
  - Efficiency
- Related topics: domain adaptation/transfer learning/few-shot learning
- Methods: Fine-tuning and its variants; Mixture of experts; Task-vectors; Matching-based methods

# Evaluation

(small T)	Fine-tuning	MoE	Task vectors	Matching
Accuracy	✓✓	✓✓✓	✓	✓
Robustness	✓	✓✓	✓✓✓	✓✓
Efficiency			Online	Online
Adapt	✓	✓	✓✓✓	✓✓✓
Test	✓✓✓	✓✓✓	✓✓✓	✓
Others		✓ Difficult to train		

# LLM-era

- In-context learning (ICL): Adaptation is just a forward pass
- Why does it work?
  - H1: Transformers implement gradient descent algorithm over IC examples
  - H2: IC examples recognize tasks in pre-training, e.g. via **task vectors**
  - H3: Self-attention implements **matching-based** adaptation via induction heads

# Re-evaluation

(small T)	Fine-tuning		ICL
Accuracy	✓✓		✓✓
Robustness	✓		✓
Efficiency			Online
Adapt	✓		✓✓✓
Test	✓✓✓		✓✓
Others			✓✓✓ Ease of use in multi-tenancy models

Improvement:

- Structuring of prompts
- Pre-training strategies
- Fine-tuning
- Example retrieval
- Model architectures

# Can LLMs Really Reason and Plan?

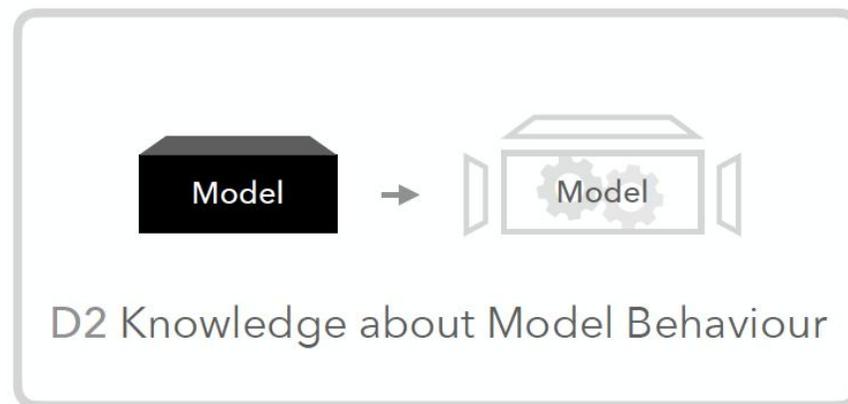
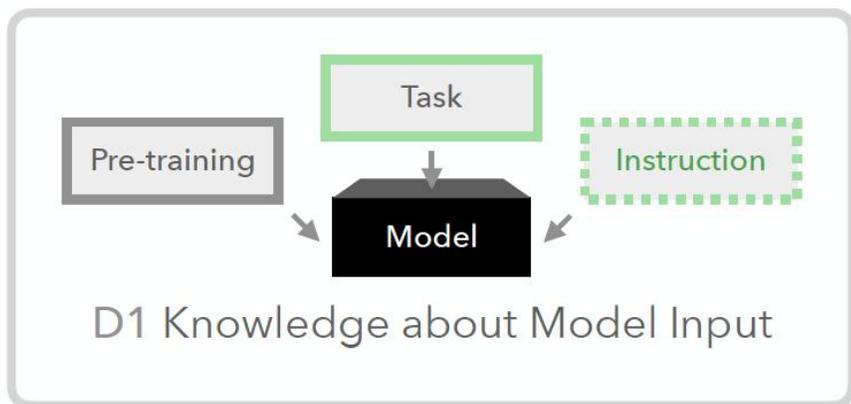


- Speaker: Subbarao Kambhampati
- Arizona State University
- 关注的问题：LLMs在推理过程中会出现幻觉等问题
- 作者对LLM常用的微调和提示语进行了很多测试
- 结论：LLMs的训练和使用方式并没有显示出它们能够进行原则性推理的迹象。

# Are LLMs Narrowing Our Horizon? Let's Embrace Variation in NLP!

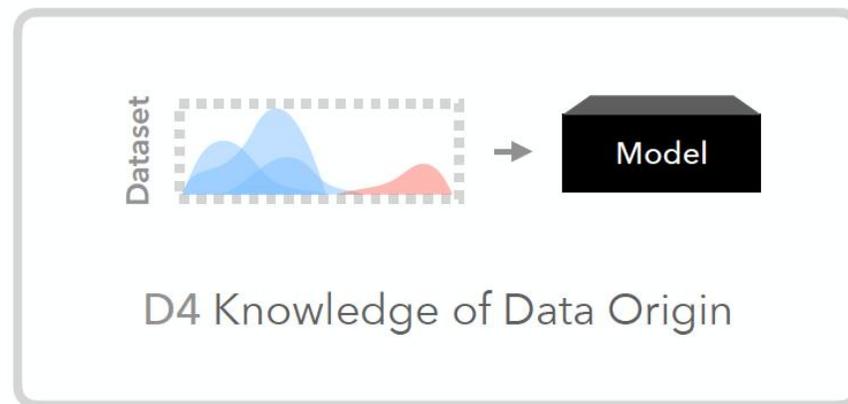
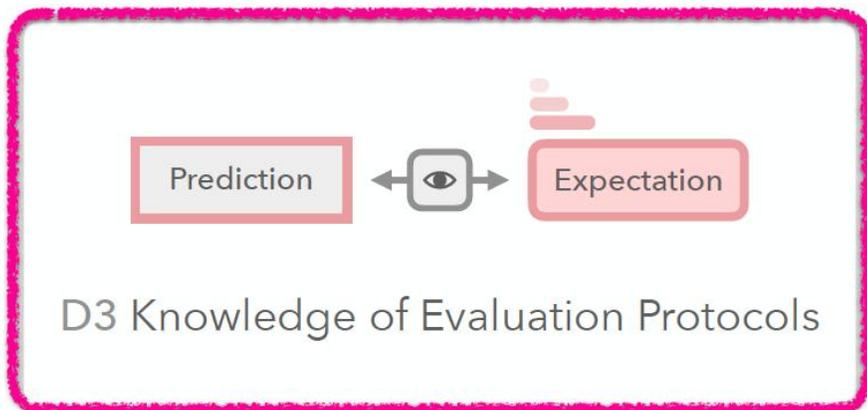
- Speaker: Barbara Plank LMU Munich & IT University of Copenhagen
- LLMs & Trust: A Short Look at AI history
- Regrain trust from four aspects:





Trust arises from knowledge of origin as well as from knowledge of functional capacity.

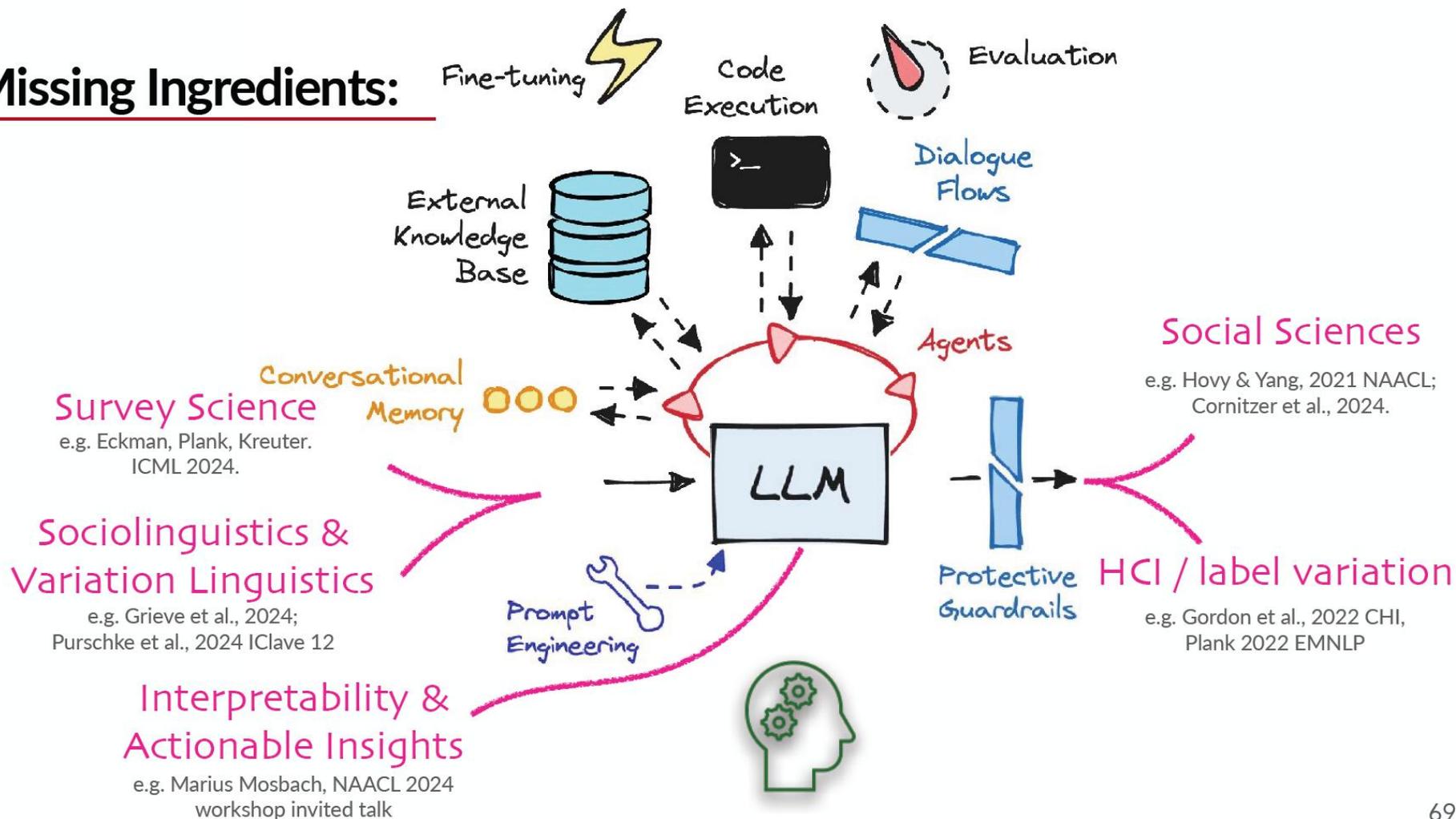
*Trustworthiness - Working Definition by David G. Hays, 1979*



# We need to embrace variation holistically

- Inputs: linguistic variation, low-resource languages & dialects
- Outputs: human label variation as signal (not error) [Uncertainty]
- Research: Language as the Bridge

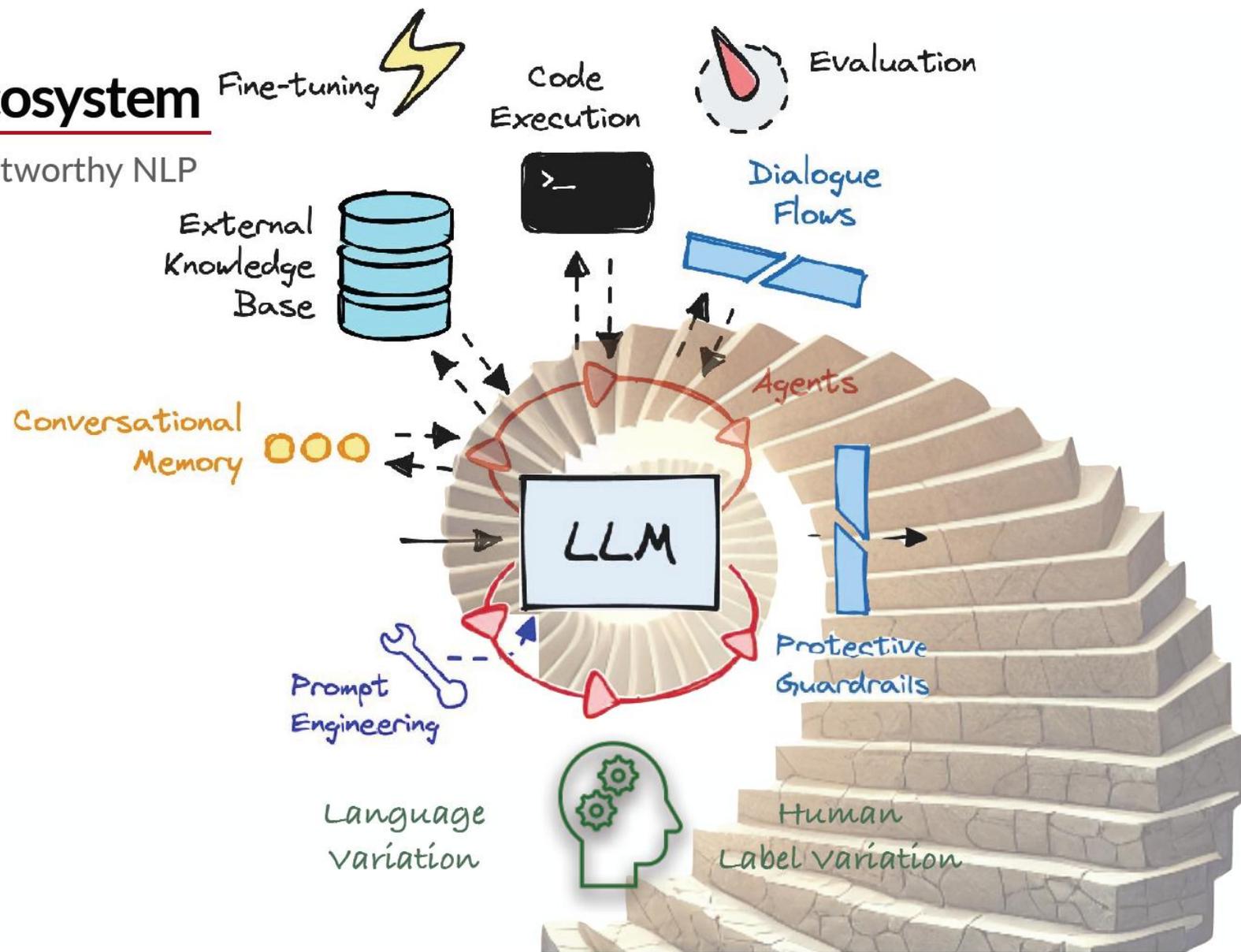
# Missing Ingredients:



69

# Trust LLM Ecosystem

Human-facing, Trustworthy NLP



# Tutorials

- **Tutorial 1– Computational Linguistics for Brain Encoding and Decoding: Principles, Practices and Beyond**

Room : Lotus 1-4 (Level 22)

*Jingyuan Sun, and Shaonan Wang, and Zijiao Chen, and Jixing Li, and Marie-Francine Moens*

- **Tutorial 2 - Automatic and Human-AI Interactive Text Generation (with a focus on Text Simplification and Revision)**

Room : Lotus 5-7 (Level 22)

*Yao Dou, and Philippe Laban, and Claire Gardent, and Wei Xu*

- **Tutorial 3 - Vulnerabilities of Large Language Models to Adversarial Attacks**

Room : World Ballroom B (Level 23)

*Yu Fu, Erfan Shayegan, and Md. Mamun Al Abdullah, and Pedram Zaree, and Nael Abu-Ghazaleh, and Yue Dong*

# Tutorials

14:00 - 17:30

Tutorial 4 - 6

- **Tutorial 4 – Computational Expressivity of Neural Language Models**

Room: Lotus 1-4 (Level 22)

*Alexandra Butoi and Ryan Cotterell and Anej Svete*

- **Tutorial 5– Watermarking for Large Language Model**

Room: Lotus 5-7 (Level 22)

*Xuandong Zhao, and Yu-Xiang Wang, and Lei Li*

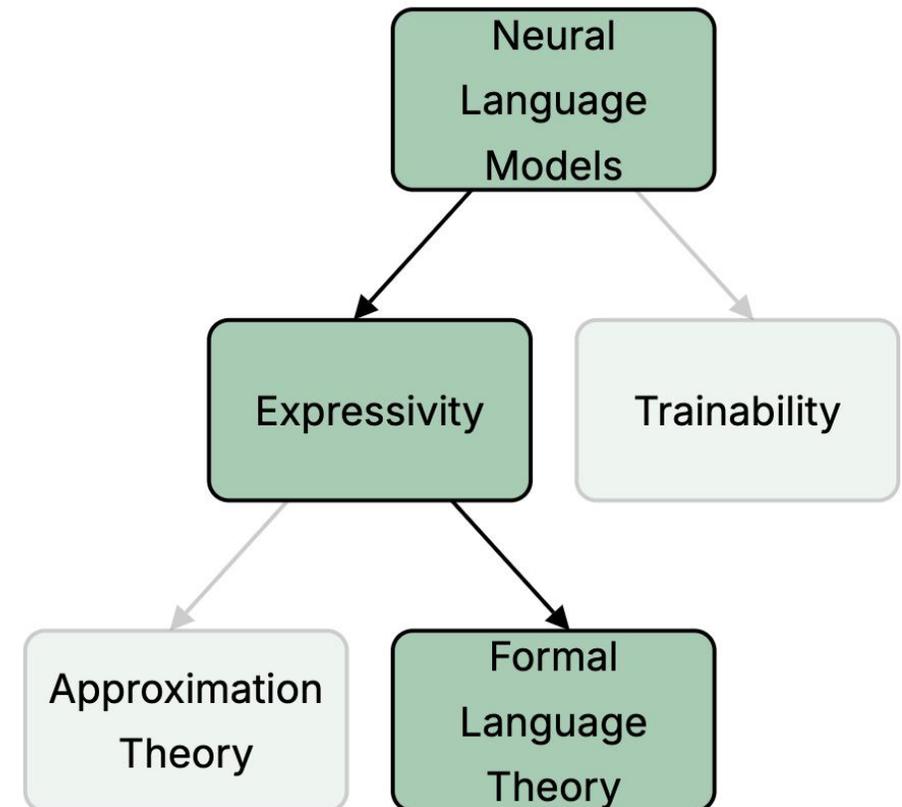
- **Tutorial 6; Presentation Matters: How to Communicate Science in the NLP Venues and in the Wild?**

Room: World Ballroom B (Level 23)

*Sarvnaz Karimi, and Cecile Paris, and Gholamreza Haffari*

# T4: Representational Capacity of Neural Language Models

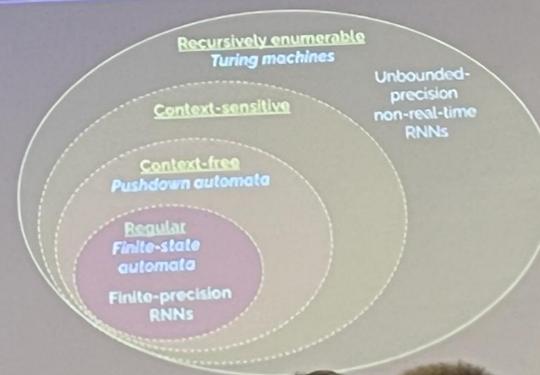
- A group of ETH Zürich led by Ryan Cotterell
- formal language theory to understand the representational capacity of LLMs
- RNN vs. Finite-State Automata
- LLMs vs. finite-state automata and Turing machines.



## Our Second Theorem of the Day

- The intuitive connection between RNNs and finite-state automata can be very concretely formalized

Theorem: Finite-precision real-time RNNs are equivalent to FSAs.



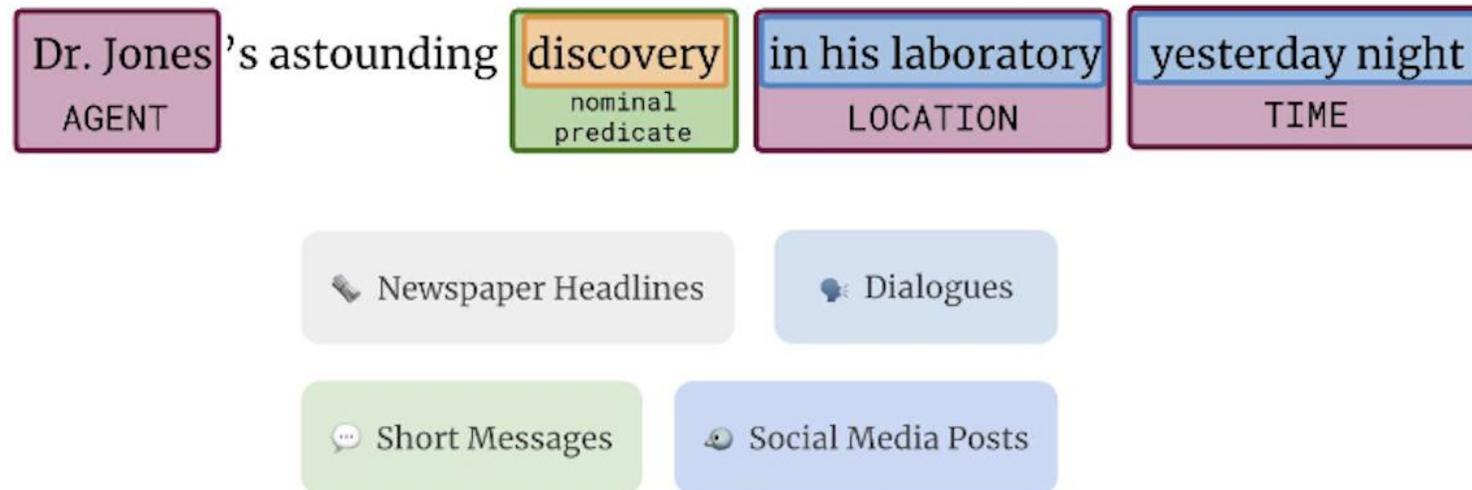
# Semantics主题 (oral)

- - NounAtlas: Filling the Gap in Nominal Semantic Role Labeling  
Roberto Navigli, Marco Lo Pinto, Pasquale Silvestri, Dennis Rotondi, Simone Ciciliano, Alessandro Scirè
- - UG-schematic Annotation for Event Nominals: A Case Study in Mandarin Chinese  
Wenxi Li | Yutong Zhang | Guy Emerson | Weiwei Sun  
Computational Linguistics, Volume 50, Issue 2 - June 2023
- Distributional Inclusion Hypothesis and Quantifications: Probing for Hypernymy in Functional Distributional Semantics  
Chun Hei Lo, Wai Lam, Hong Cheng, Guy Emerson

# NounAtlas: Filling the Gap in Nominal Semantic Role Labeling

## Nominal SRL is under studied

Existing SRL has been primarily focused on verbal predicates. But nominal predicates are frequent in real-world settings!



# Inventory

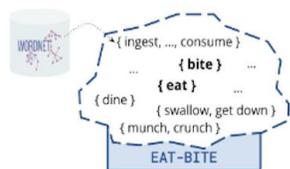
## NounAtlas

Filling the gap in nominal SRL with NounAtlas!

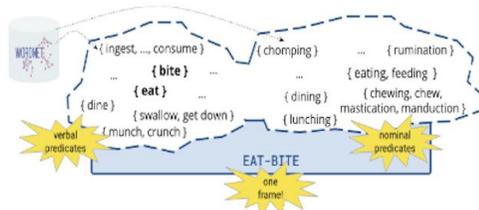
### Creating a huge inventory for Nominal SRL

#### Expanding VerbAtlas frames

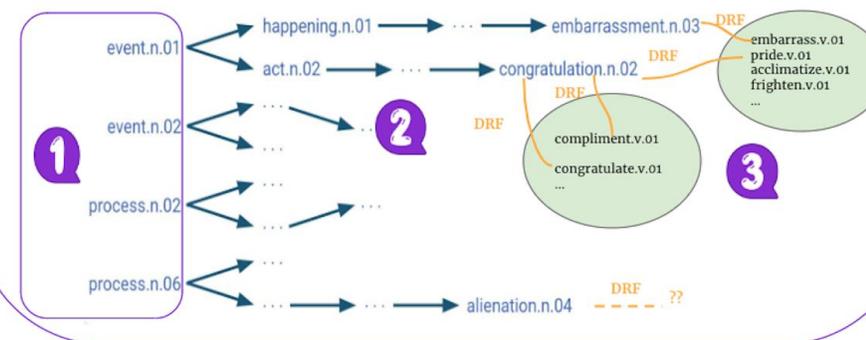
We start from the existing VerbAtlas frames, which groups together semantically-related verbal predicates.



We expand the VerbAtlas frames by adding WordNet nominal synsets that convey similar semantics.



- 1 We manually selected the most prominent **nominal synsets from WordNet**
- 2 We collect all their **nominal descendants** (i.e. via direct/indirect WordNet **hyponym** relations)
- 3 We **link** nominal synsets to their verbal counterparts through WordNet's **Derivationally Related Forms (DRF)** edges

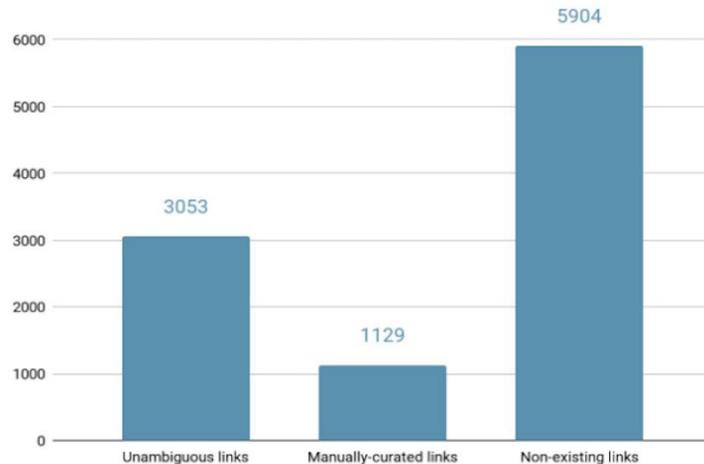


# Evaluation

## Wordnet-based synset-to-frame mapping: Evaluation

1. **Unambiguous links**, evaluated on 100 items: **82% in agreement with manual annotations**, 18% made up of equally-valid selections

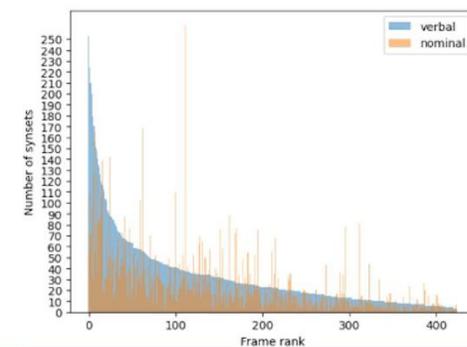
2. **Ambiguous links**, evaluated on 100 items: tie broken through manual curation, Cohen's Kappa = 0.57 (moderate)



3. **Missing links**: We use a Cross-Encoder trained on the other types of link to score predicates and, aggregated on  $\leq 10$  predicates, frames. We consider the top-5 frames and manually link.



- As a result of manual annotation, **99.6% of Unlinked synsets were added to a frame**
- **88.2% of correct frames contained in our model's top-5 ranking**, 49.3% in the top-1



# Corpus + model

## Creating a nominal SRL corpus

### Predicate nominalization: Generation

Idea → start from SemCor (Miller et al. 1993):  
Verbal predicates in SemCor annotated with WordNet synsets are directly linked to VerbAtlas frames!

- We retain all SemCor sentences featuring verbal predicates
- We generate the corresponding nominalized version of each sentence using Gemini-Pro

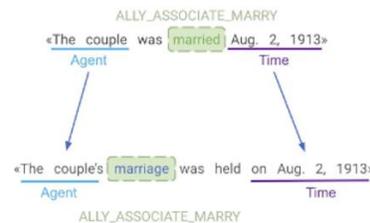
Change the sentence by nominalizing the verb "married" indicated by \*\*. Use exactly one of these deverbal nouns: "marriage ceremony", "marriage", "wedding". Indicate the chosen deverbal noun with \*\*: "The couple was \*\*married\*\* Aug. 2, 1913."  
"The couple was ""marriage"" Aug. 2, 1913."

✦ The couple's ""marriage"" was held on Aug. 2, 1913.

- We validate the generation by keeping the sentence if:
  1. The deverbal noun is **identified**
  2. The deverbal noun is a **noun**
  3. Its **lemma is among the candidates** provided

### Verbal-to-nominal role propagation

We annotate the verbal sentences with InVeRo-XL (Conia et al. 2021)



We validate ~500 random sentences to generate a gold test set for Nominal SRL:

unchanged sentences: 92.11%  
unchanged frames: 95.56%  
unchanged role spans: 77.11%  
unchanged role labels: 95.27%



## A unified model for SRL

### Setup

Backbone model: RoBERTa-based model from Conia and Navigli (2020)

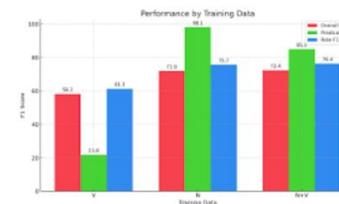
Training data:

- Nominal SRL: Our dataset
- Verbal SRL: OntoNotes 5.0 (V)
- The combination of the two datasets (N+V)

Test data:

- Our ~500 manually-curated nominal sentences (nouns)
- The OntoNotes test set (verbs) converted to VerbAtlas annotations

Gold-standard nominal test set



OntoNotes verbal test set



Our model, jointly-trained on nominal and verbal data (N+V), achieves competitive performance on nominal SRL!

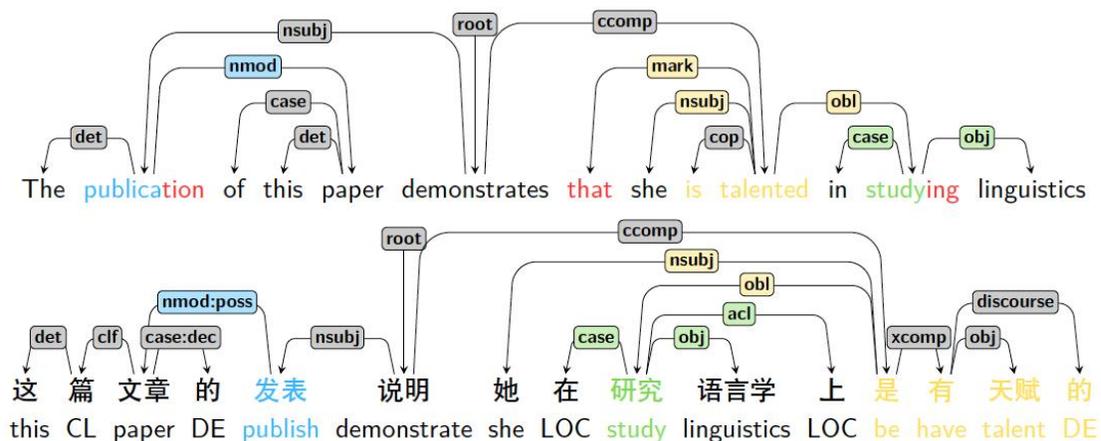


Also robust on the OntoNotes test set (verb frames).

Thanks to NounAtlas, we devised a unified approach for nominal and verbal SRL!

# Background

## Multilingual Heterogeneity of Event Nominals



- **Identification:** Languages differ in their use of functional morphemes for nominalizations and the types of morphemes they use.
- **Nominal SRL:** Languages differ in how they realize participants of event nominals through syntactic constructions

# Research Question

## Multilingual Heterogeneity of Event Nominals

	Mandarin	English	German
Subject with nominative case	–	?	+
Accusative case on object	–	?	+
Projection of outer Aspect	–	+	+
Modal or auxiliary verb	+	+	+
Complementizer	–	+	+
Verb suffix	–	+	+
Genitive/PP-subject	–	+	+
Genitive/PP-object	–	+	+
Gender	–	?	+
Quantity	–	+	+
Determiner	–	+	+
Noun suffix	–	+	+

Table: Verb- and noun-related features of event nominals

### Research Question

**How can we accommodate multilingually heterogeneous phenomena in a unified way, or more theoretically, uncover universals of the world's languages beneath their surface-level variations?**

# Distributional Inclusion Hypothesis and Quantifications: Probing for Hypernymy in Functional Distributional Semantics

## DIH and Quantifications

DIH.  $r_2$  is a hypernym of  $r_1$  iff  $r_1$ 's characteristic contexts  $\subseteq r_2$ 's.

Quantifications. A corpus with only universally quantified statements results in the reverse of DIH (rDIH).

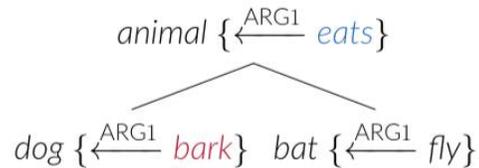


Figure 1. A taxonomic hierarchy of nouns. Next to each noun are the contexts applicable to it and its descendants.

### Corpus 1 (DIH)

some dog *barks*  
 some animal *barks*  
 some bat *flies*  
 some animal *flies*  
 some animal *eats*

### Corpus 2 (rDIH)

every dog *barks*  
 every dog *eats*  
 every bat *flies*  
 every bat *eats*  
 every animal *eats*

Table 1. Corpora generated from the hierarchy in Fig. 1.

## FDS

Entity Vectors.  $z \in \mathbb{R}^d$

Truth-Conditional Semantic Functions.

$$t^{(dog)}(z) = P(dog(z) = \top | z) \\ = \text{sigmoid} \left( v^{(dog)\top} z + b^{(dog)} \right)$$

Representing Hypernymy.

$$\forall z \text{ s.t. } \|z\|_2 \leq 1: t^{(dog)}(z) < t^{(animal)}(z)$$

which is true iff  $s(dog, animal) > 0$ , where

$$s(r_h, r_H) = b^{(r_H)} - b^{(r_h)} - \left\| v^{(r_H)} - v^{(r_h)} \right\|_2$$

Model Training. By Lo et al. (2023), given a DMRS graph:

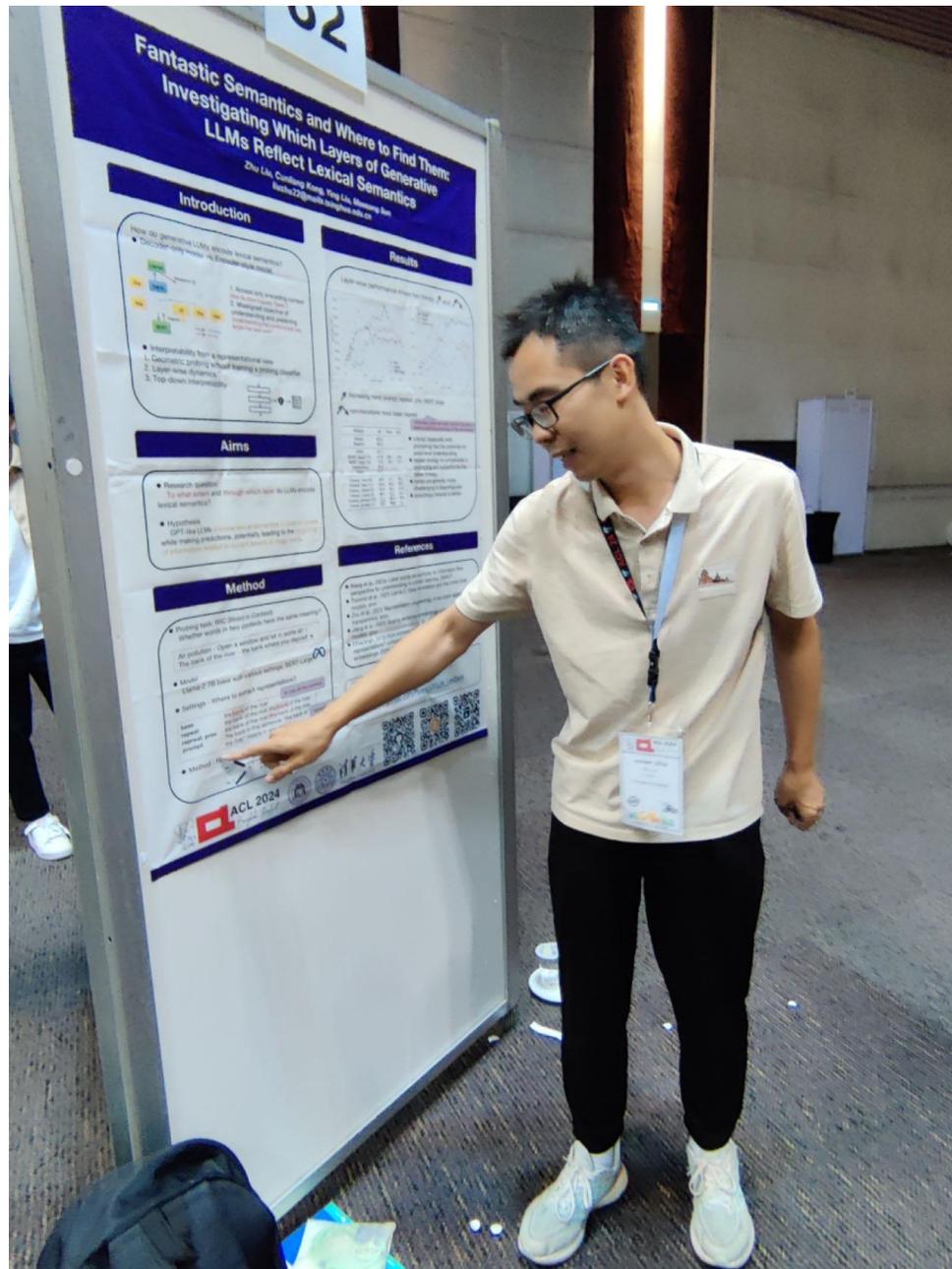
$$\text{some} \xrightarrow{\text{RSTR}} \text{dog} \xleftarrow{\text{ARG1}} \text{bark}$$

Variational Inference:  $q_\phi(z | \xleftarrow{\text{ARG1}} \text{bark})$

Reconstruction:  $\max \ln \mathbb{E}_{q_\phi(z | \xleftarrow{\text{ARG1}} \text{bark})} [t^{(dog)}(z)] + \dots$

**New objective (FDS<sub>v</sub>).** Optimizes over regions of the entity space for handling universal quantifications

# Posters



# What does Kiki look like?

## Cross-modal associations between speech sounds and visual shapes in vision-and-language models

Tessa Verhoef, Kiana Shahrasbi and Tom Kouwenhoven

Leiden Institute of Advanced Computer Science, Leiden University, The Netherlands  
Correspondence: t.verhoef@iacs.leidenuniv.nl, t.kouwenhoven@iacs.leidenuniv.nl



### INTRODUCTION

Humans have clear cross-modal preferences when matching novel words to visual shapes.

Example: **bouba-kiki effect!**

The development of multimodal models, including VLMs, can potentially revolutionise how machines understand and interact with humans.

**Do VLMs associate non-words and visual stimuli in a human-like way?**



Bouba or Kiki? From [5,9]

### BACKGROUND

Non-arbitrariness as a **general property of language** [8]. Affects language **learning** and shapes language **emergence** [4,9,10].

**Possible sources:**

- o Orthography [2]
- o Acoustics and articulation [6,9,11]
- o Affective-semantic properties of vocal communication [7]
- o Physical properties relating to audiovisual regularities in the environment [3]

Alper and Averbuch-Elor [1] reported strong evidence for a bouba-kiki effect in CLIP and Stable Diffusion's surprising given model training and the absence of relevant data sources such as auditory information and experience with physical object properties.

### METHODS

**Models Architecture Attention #Params #img,caps (M)**

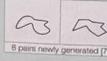
Model	Architecture	Attention	#Params	#img,caps (M)
VLT	Single-Stream	Multi-head	87.4M	4, 10, 30, 90
BLIP2	Dual-Stream	Q-Former	3.8B	128, 256
GPT-4o	Transformer	Multi-head	11.7B	128, 256
Llama2	Transformer	Multi-head	70B	128, 256

**Experiments:** Probed VLMs using a cognitive science paradigm to test for the bouba-kiki effect.

**Images:**

Originals 0 from [5,9] as well as other image pairs from prior human experiments and entirely novel generated images, following [7].

Examples ⇨

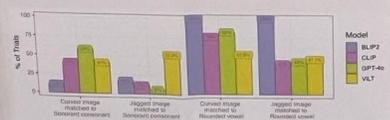


**Pseudowords:** from [7]

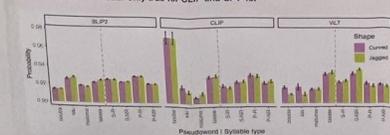
Sonorant consonants:	Plosive consonants:
mi, ni	k, k, p
Rounded vowels:	Non-rounded vowels:
oo, oo, uh	ee, uh, ay
<b>Examples:</b>	<b>Examples:</b>
moonch, loonah, nahmo	teekay, kuhpab, teekay

**4 separate tests:** Pseudoword generation (PWG) and probability score comparison (PSC), each with single syllables (1) or two-syllable words (2).

### RESULTS



**Fig 1:** Curved or jagged images matched to Sonorant consonants (left bars) or Rounded vowels (right) in **PWG1** test. The expected pattern would show higher bars for the Curved than for the Jagged shapes in both sets. Only true for CLIP and GPT-4o.



**Fig 2:** Probability scores in **PSC1** for two pairs of original pseudowords (bouba & kiki, taekoo & moonmo) = four generated syllable types. Respondent (rounded (R) or jagged (J)) expected to be most Curved, Sonorant (non-Rounded (S) or Plosive (Rounded (R)) and Plosive (non-Rounded (P)) V, expected to be most Jagged, paired with Jagged or Curved shapes. No effect found.

**Fig 3:** Evidence for a bouba-kiki effect found in only three tests (ignoring CLIP and GPT-4o. VLT and BLIP2 never passed)

### DISCUSSION

It is too early to conclude that VLMs understand sound-symbolism or map visio-linguistic representations in a human-like way. Results depend heavily on which specific model is tested and how the task is formulated.

However, results tentatively suggests that cross-modal preferences can, to some extent, be learned from statistical regularities in data. Model features such as architecture design, training objective, number of parameters, and input data seem to affect the results.

**These findings inform discussions on the origins of the bouba-kiki effect in human cognition and future developments of VLMs that align well with human cross-modal associations.**

### REFERENCES

- Alper, E., & Averbuch-Elor, E. (2023). The bouba-kiki effect in CLIP and Stable Diffusion. *arXiv preprint arXiv:2305.18020*.
- Alper, E., & Averbuch-Elor, E. (2023). The bouba-kiki effect in CLIP and Stable Diffusion. *arXiv preprint arXiv:2305.18020*.
- Alper, E., & Averbuch-Elor, E. (2023). The bouba-kiki effect in CLIP and Stable Diffusion. *arXiv preprint arXiv:2305.18020*.
- Alper, E., & Averbuch-Elor, E. (2023). The bouba-kiki effect in CLIP and Stable Diffusion. *arXiv preprint arXiv:2305.18020*.
- Alper, E., & Averbuch-Elor, E. (2023). The bouba-kiki effect in CLIP and Stable Diffusion. *arXiv preprint arXiv:2305.18020*.

# Aligning to Adults Is Easy, Aligning to Children Is Hard: A Study of Linguistic Alignment in Dialogue Systems

Dorothea French, Sidney D'Mello, Katharina von der Wense  
University of Colorado Boulder

### Introduction

Over the course of a conversation people start to align with each other. They use similar words, syntax, and talk about similar ideas. This phenomenon, called linguistic alignment, helps conversational participants understand each other and reduces the effort needed to do so, amongst other positive effects.

Linguistic alignment is particularly important when one of the participants is a child, or a non-fluent speaker of the language. Levels of parent's alignment have been shown to correspond to a child's vocabulary and language development. Repetition and correction also help language learners improve naturally.

In this era, characterized by widespread use of dialogue systems, it is important to ask how well they can communicate with humans. Alignment of dialogue systems can reduce frustration and perceived task difficulty. It is also a critical component of communication. Thus we must ask: how well do our systems do? Do they reach human-like levels, or is this an area of friction?

In this work we ask two questions:

**To what degree do two state-of-the-art dialogue systems, ChatGPT and Llama2 - align to users, and how does their alignment compare to that typically seen between humans?**

**To what degree do ChatGPT and Llama2 align to children (i.e. non-fluent speakers), and how does this level of alignment compare to a parent's?**



### Approach and Data

We use the Switchboard Dialogue Acts Corpus (SWDA) of a series of phone conversations on a variety of topics to assess model alignment with adults.

**SWDA Context**

A	Any jury's not going to disregard the evidence, you know.
B	Uh, that's true.
B	I, I, I think our judicial system is attorney welfare myself!
A	That may very well be.
B	I hold it in the utmost contempt.
GPT	it's definitely a possibility that needs to be looked into.
Llama2	Yeah, it's like, you know, they're just trying.

Table 1: Last lines of SWDA transcript plus true and generated responses.

From the datasets we select transcript excerpts of 36 turns - 35 turns of context, and to following utterance.

Our two models, GPT-3.5 and Llama2, are prompted with respond to the 35 turns of context as a participant or parent respectively.

We use the CHILDES dataset of conversations between caretakers and children to assess the models' alignment to children, i.e. language learners.

**CHILDES Context**

MOT	Hm?
CHI	Where Mommy go?
MOT	Mommy went to the university this morning to get some books.
CHI	Where's Mommy's books?
Response	
MOT	They're in the hallway in a big bag.
GPT	Mommy will bring the books home this evening.
Llama2	Mommy left her books in the car.

Table 2: Last lines of CHILDES transcript plus true and generated responses.

We look at three types of alignment (syntactic, lexical, and semantic) between the last turn of the context and the following utterance.

We compare these metrics between the models and gold responses.

### Results

**Table 3: Average alignment for responses to SWDA.**

Response	Syntactic	Lexical	Semantic
True	0.444	0.170	0.308
ChatGPT	0.443	0.151	0.340
Llama2 13b	0.472	0.207	0.350
Llama2 7b	0.475	0.213	0.374

Table 3: Average alignment for responses to SWDA.

**Table 4: Average alignment for responses to CHILDES.**

Response	Syntactic	Lexical	Semantic
True	0.490	0.278	0.411
ChatGPT	0.438	0.190	0.347
Llama2 13b	0.464	0.227	0.345
Llama2 7b	0.473	0.251	0.370

Table 4: Average alignment for responses to CHILDES.

Llama2 is more repetitive, and it does a very good job mimicking the stylistic elements of the conversation, but often makes less sense.

When responding to children, both models have lower than human levels of alignment. Like with adults, ChatGPT has a slightly better proportion of very poor (1-2 out of 5) responses. Once again, Llama2 has more repetitive responses, but higher alignment, in both response sets we see the smaller Llama2 model consistently has higher alignment.

### Introduction

Linguistic alignment is a critical component of dialogue. ChatGPT, does a fairly good job of aligning at human levels during normal conversation. Llama2 had higher alignment, but decreased response quality. Llama2's responses are more "parrot-like", and it does not introduce novel information. This shows alignment beyond human-levels is not, alone, a sufficient metric to judge the performance of a dialogue system. In fact, beyond that point it negatively correlates with quality.

However, linguistic alignment to children or language learners plays an additional role to support learning. As such, the elevated alignment exhibited by Llama2 may have benefits, such as help develop the child's language skills. This type of use might also care less about novelty and helpfulness of the system, and more about ease of understanding and reduction in cognitive load. Thus, alignment metric when working with educational dialogue systems.

### Conclusion

Dialogue systems show great potential to assist humans across a variety of tasks. The success of these interactions with linguistic correlates when responding to adult speakers. ChatGPT shows approximately human-level alignments and provides constructive responses. Llama2, however, over-mimics the conversation, this could be positive when talking with children or language learners. It results in elevated alignment. We conclude that SOTA dialogue systems can improve in regards to tailoring levels of alignment to match various circumstances, without reducing quality. In the future, we plan to investigate alignment to adult learners, non-typical speakers, and explore methods to create dialogue systems with a closer to human levels of alignment.



# Space Decomposition for Sentence Embedding

Wuttikorn Ponwitayarat\*, Peerat Limkonchotiwat†, Ekapol Chuangsuwanich, Sarana Nutanong†

\* School of Information Science and Technology, VISTEC, Thailand  
† Department of Computer Engineering, Chulalongkorn University, Thailand

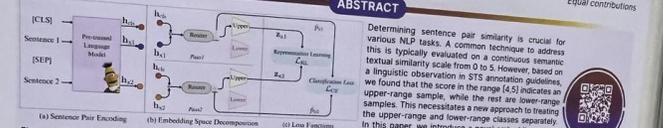


Figure 1: The overview of Mixture of Specialized Projectors (MxSP). (a) Given an upper-range sample, we encode the sample with a pre-trained language model. We use a router to classify a class of sentences 1 and 2 (upper-range or lower-range). The final representation is formulated by projecting the representation with specialized projectors. (c) We improve the classification and representation with our training losses  $\mathcal{L}_{CLS}$  and  $\mathcal{L}_{CLS}$ , respectively.

- CONTRIBUTIONS**
- We have recast the sentence embedding paradigm from one embedding space containing upper-range and lower-range to separate embedding space for each group.
  - We propose a novel embedding space decomposition technique called **Mixture of Specialized Projectors (MxSP)**. Our model has the ability to distinguish upper-range and lower-range samples while accurately ranking sentence pairs within each class.
  - We demonstrate the efficiency of our method on STS and zero-shot benchmarks. In addition, we provide deep analysis of (I) Performance efficiency and (II) Design choice in embedding space decomposition settings.
- How do we decompose the embedding space?**
- We formulate inputs in cross-encoder format:  $\{CLS, w_{u1}, w_{l1}, w_{u2}, w_{l2}\}$
  - We separate  $w_{u1}$  and  $w_{l1}$  into upper and lower ranges using the STS score where the lower range is  $\{0, 4\}$  and the upper range is  $\{4, 5\}$
  - We then optimize the classification head using binary cross-entropy:  $\mathcal{L}_{CLS} = -\sum_{i=1}^n \log(\sigma(\mathcal{F}(w_i)))$
  - Finally, we send upper and lower samples to separate (specialized) projectors:  $R_u = \{r_{u1}, \dots, r_{un}\}$ ,  $R_l = \{r_{l1}, \dots, r_{ln}\}$

## EXPERIMENTAL RESULTS

Table 1: Spearman's rank correlation on the STS benchmarks.

Method	STSB	STSC	STSD	STSE	STSF	STSG	STSH	STSI	STSL	STSM	STSN	STSO	STSP	STSQ	STS
BERT	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88
MxSP	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89

Table 2: The MAP score on the reranking task. Table 3: The AUC score of binary task classification on news standard datasets.

Table 4: The design choice of our framework. We evaluate the average STS score across news STS datasets - in replicating the test method with the right method.

Method	STSB	STSC	STSD	STSE	STSF	STSG	STSH	STSI	STSL	STSM	STSN	STSO	STSP	STSQ	STS
BERT	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88
MxSP	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89

# Transformer Attention vs Human Attention in Anaphora Resolution

Anastasia Kozlova<sup>1</sup>, Albina Akhmetgareva<sup>1</sup>, Aigul Khanova<sup>2</sup>, Semen Kudriatsev<sup>2</sup>, Alena Fenogenova<sup>1</sup>  
<sup>1</sup>SaluteDevices, <sup>2</sup>HSE University

- Contributions**
- EyeWino - a new dataset based on human eye-tracking data during anaphora resolution;
  - A set of experiments on different models fine-tuned on the data to explore the attention mechanisms;
  - An extensive analysis of the correlations between human and machine attention maps;
  - Experiments with the human gaze integration into the transformer's attention mechanisms.
- Objective**
- investigate the correlation between human and machine attention for anaphora resolution;
  - improve models for anaphora resolution using insights from human attention patterns.
- Methodology**
- WinoGram Schema (WS) task is devoted to anaphora resolution in the specifically designed experiment, where reference could be resolved only using common sense.
- Task format**
- Human: "Bob collapsed on the sidewalk. Soon he saw Carl coming to help. HE was very concerned. Does the highlighted pronoun refer to Carl / Bob?"
- Human Attention**
- Total reading time: the sum of all fixation durations on the current word, ms
  - Gaze duration: the sum of all fixation durations on the current word in the first pass reading, ms

**Transformer Attention**

We extract the attention weights from the encoder layers and average them across all attention heads. The matrix aggregations to obtain a vector of token importance:

- mean - the average of all the rows in each column;
- row - the average of pronoun tokens in each column.

**Models**

- BERT-based models include ruBERT-base and mBERT-base;
- ruBERTa-based models include ruBERTa-large and XLM-R-large;
- TS-based models include ruTS-base and mTS-base.

**Fine-tuning datasets**

- TAPE, the Russian WS Challenge dataset;
- MERA, the Russian WS Dataset from the MERA benchmark;
- XWINO, the multilingual WS Challenge dataset.

**Results**

**Spearman Correlation**

Model	App	Checkpoint	Layer	Fixations	Gaze duration	Total reading time	Without word
ruBERT-base	mean	pre-trained	1	0.601	0.302	0.500	0.339
mBERT-base	mean	pre-trained	1	0.603	0.304	0.500	0.340

**Accuracy**

Model	App	Checkpoint	Layer	Fixations	Gaze duration	Total reading time	Without word
ruBERT-base	mean	pre-trained	1	0.772	0.470	0.711	0.470
mBERT-base	mean	pre-trained	1	0.773	0.471	0.712	0.471

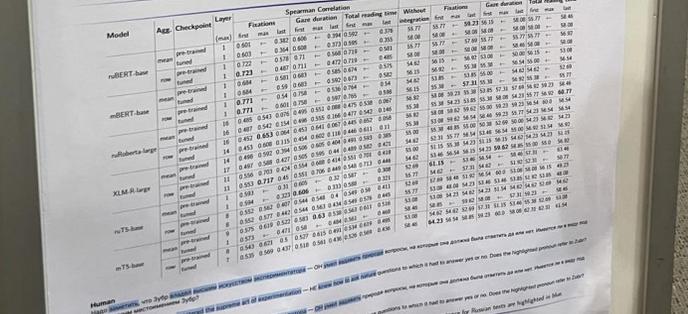


Figure 1: Visualizations of human and most correlated model's attention. The words with high relative importance for human focus are highlighted in blue. It should be noted that zero (0) means the attention is 0.0. The blue bars indicate the attention weights on the highlighted words. The model's attention is more fact-oriented than human's attention. The model's attention is more fact-oriented than human's attention. The model's attention is more fact-oriented than human's attention.

# Most hallucinations are wrong, but some are useful. We are calling them confabulations — a useful alignment with the human behavior of storytelling.

Confabulation: The Surprising Value of LLM Hallucinations  
Peiqi Sui, Eamon Duede, Sophie Wu, and Richard Jean So

- 1 Motivation**
- Despite research suggesting the potential inevitability of AI hallucinations, computer scientists are thinking about them too narrowly, guided by an impossible ambition to eliminate all hallucinations and dubious normative commitment to sanitize any perceived risk of AI. We instead argue for a broadened view of hallucinations in LLMs as "confabulations," valuable for the communicative and cognitive affordances of their higher narrative and coherence, mirroring human storytelling. This affordance-centric view provides a framework for negotiating our co-existence with AI hallucinations and challenge the view that they are purely detrimental.
- 3 Confabulations are narrative-rich and more coherent**
- In all three datasets, outputs labeled as 'hallucinated' have on average the highest narrativity



- 2 Data and Methods**
- We conduct empirical experiments on three popular hallucination benchmarks, FaithDial, BEGIN, and MERA, to validate confabulation as a storytelling and sense-making resource instead of inherent pitfall:
- Modeling Narrativity with an ELECTRA-large model finetuned on the StorySeeker dataset
  - Measuring Coherence with the DEAM model
  - Comparing the narrative and coherence scores of confabulations vs. strictly factual outputs

Variable	Coefficient	Std. Error	Variable	Coefficient	Std. Error
Narrativity	0.631***	0.059	Narrativity	0.372***	0.029
Interest	0.368***	0.038	Interest	0.433***	0.018

- Logistic regression results that demonstrate the correlation between narrativity and hallucination (left), and narrativity and coherence (right).
- 4 The Affordances of Narrativity**
- Narrative theory finds that stories enrich our cognitive landscape by making our internal world models more robust. Medical research have also validated the role of narrative interventions in improving patient care



# Social Events



# 颁奖环节

- The 2024 winner of the 1999 Test-of-Time Paper Award is:
  - Lillian Lee. 1999. Measures of Distributional Similarity.
  - In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL), pages 25–32, College Park, Maryland, USA.
- The 2024 winner of the 2014 Test-of-Time Paper Award is:
  - Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation.
  - In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, Doha, Qatar.

# Best Paper Award

- Mission: Impossible Language Models
- Why are Sensitive Functions Hard for Transformers?
- Deciphering Oracle Bone Language with Diffusion Models
- Causal Estimation of Memorisation Profiles
- Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model
- Semisupervised Neural Proto-Language Reconstruction
- Natural Language Satisfiability: Exploring the Problem Distribution and Evaluating Transformer-based Language Models

# One more thing... from EMB

## ACL is Not an AI Conference

- I am not using “AI” as a synonym for “ML”.
- Machine learning (including deep learning) provides many techniques that are useful for language technology and computational linguistics
  - (Though both of those terms are problematic.)
- The problems of CL/NLP can also be illuminating for questions about ML
  
- The issues that I am concerned with arise when the focus shifts to “AI”

## AI as a research & commercial field

Asks questions like:

- How do we build “thinking machines” that can do “human-like” reasoning?
- How do we build “thinking machines” that can “surpass” humans in cognitive work?
  - (and cure cancer, solve the climate crisis, make end-of-life decisions, etc)
- How do we automate the scientific method?
- How do we automate away such creative work as painting and writing?
  - Or: How do we steal artwork at scale and try to convince people this is “for the common good”?

# Compling/NLP asks questions such as

- How are languages similar/different?
- How is information represented in languages?
- How can we build technology that assists with: transcription, translation, summarization, information access ... in different languages?
- How can we evaluate such technology?
- What kinds of intermediate representations are useful for such technology?
- How well do different ML techniques work for different tasks?
- How do language technologies interact with existing systems of power and oppression?

## ACL is historically and should remain:

- A venue for people who care about the *language* in language technology
- A community that fosters interdisciplinarity
- A research field that cares about language communities
- ... and, as a result, a space where we can reason about societal impacts of our research and technology

<https://bit.ly/EMB-ACL24>

**Language processing is a prerequisite for “AI”, but that doesn’t mean that “AI” is the only goal of CL/NLP.**

# Workshops

- Rep4NLP (Probing -> causal inference)
  - - Latent Space Exploration for Safe and Trustworthy AI by Hassan Sajjad
  - The gap between what language models say and what they know
  - Generalisation in LLMs – and beyond
  - Efficiency as an Inductive Bias: Towards Tokenizer-free and Dynamically Sparse Language Models
- Heng Ji: AI Plays Medicinal Chemist
- Anna Rogers: Ai for reasearch workflow

# Q & A

THANK YOU

# Note