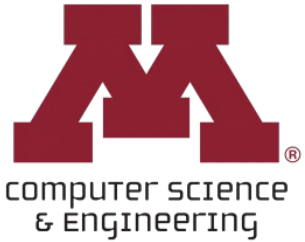
Human-centric NLP in Era of Large Language Models

Some latest work done at Minnesota NLP group

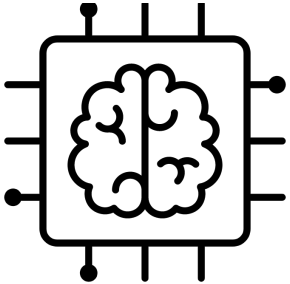
Dongyeop Kang (DK)
dongyeop@umn.edu | twitter.com/dongyeopkang | dykang.github.io

CSCI 5541 S24
20240418

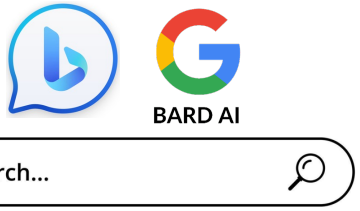


Era of Large Language Models (LLMs)

Multi-task generalization

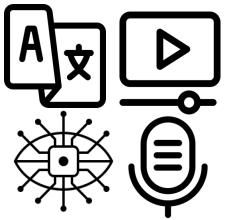


Large-language Model



Generative Search

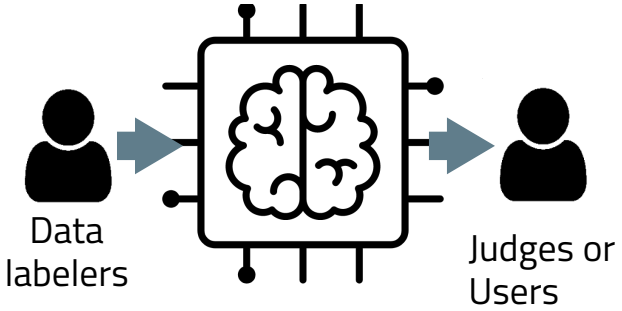
Human preferable responses



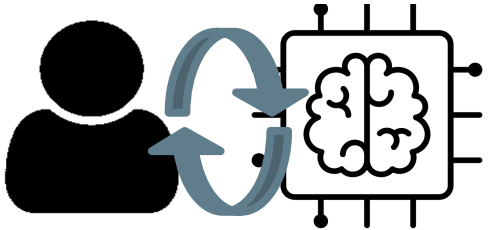
Multi-lingual, Multi-modal

Generative Agents

Tool Integration



Model-centric NLP

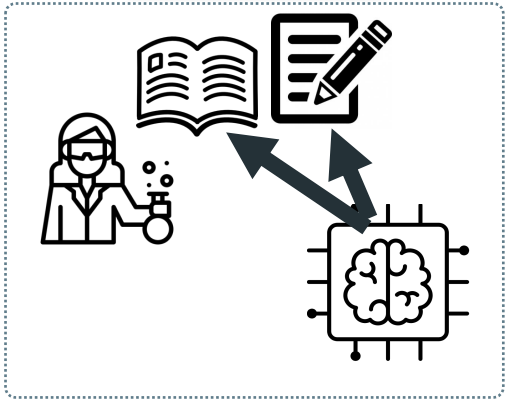


Human-centric NLP

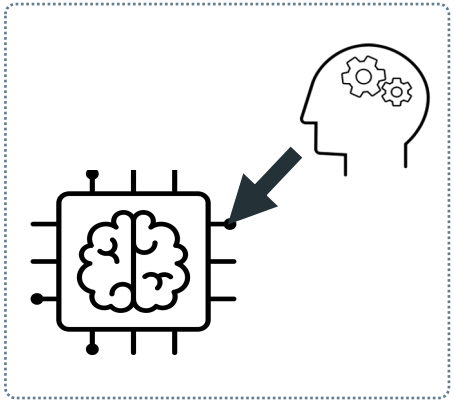
Human-centered NLP

- The advent of LLMs has reshaped the landscape of AI research, challenging traditional boundaries and raising concerns about copyright, unemployment, and ethical issues.
- Understanding and harnessing the capabilities and risks of LLMs for the benefit of human, society, and experts.
- Building more human-centric AI systems learning from human cognition, societal values, and expert skills

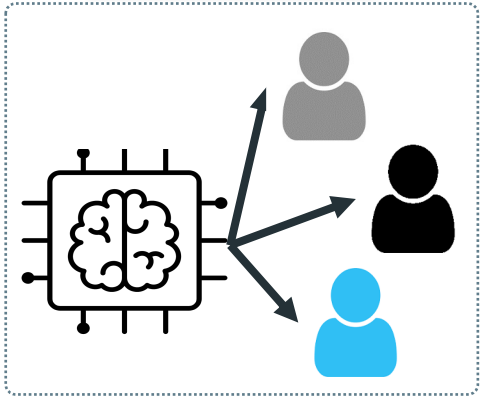
Expert-level AI

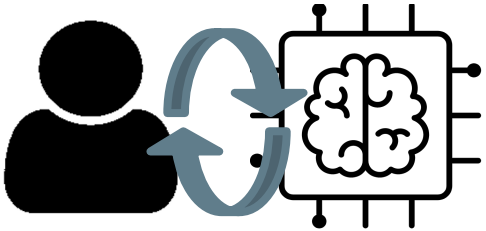


Cognitive Scaffolding

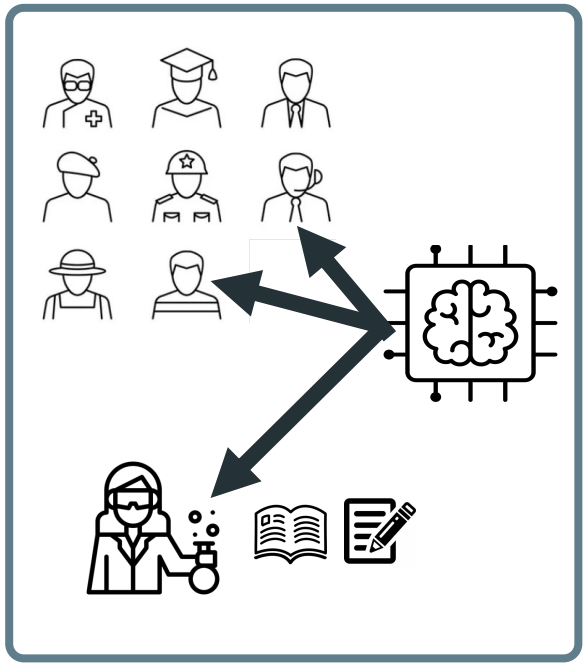


Societal Alignment



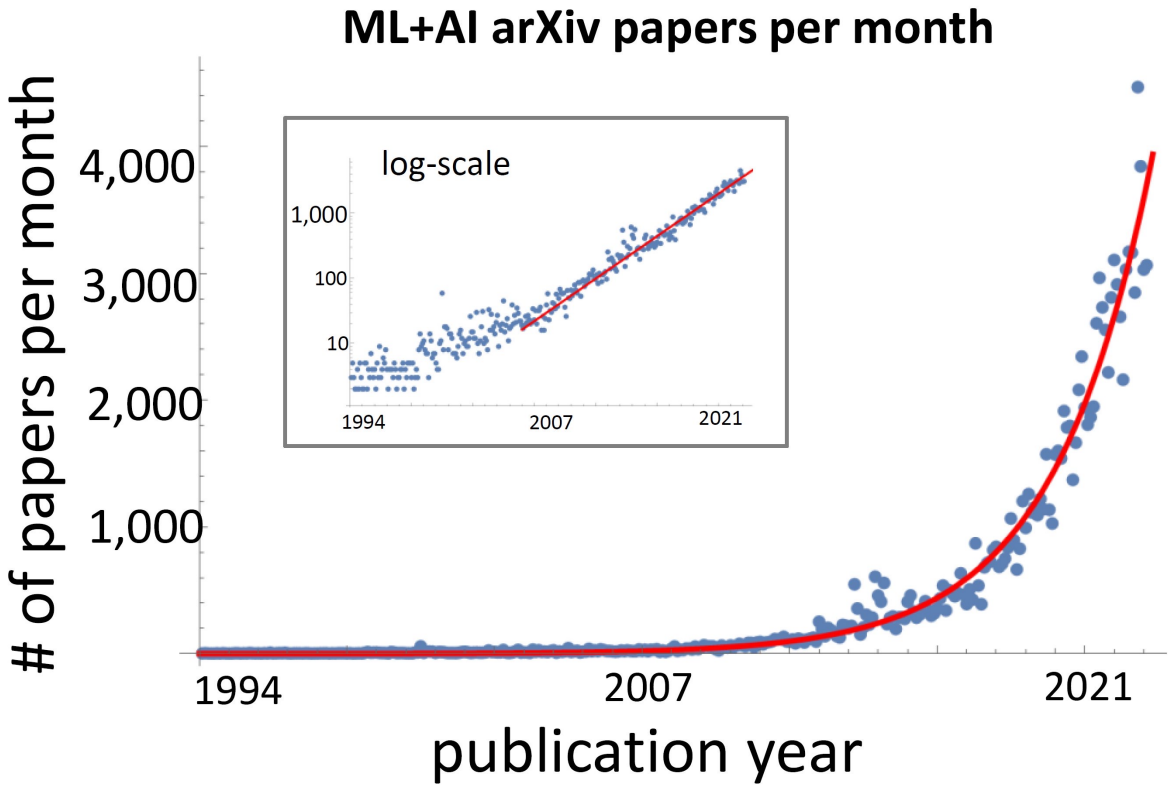


Human-centric NLP



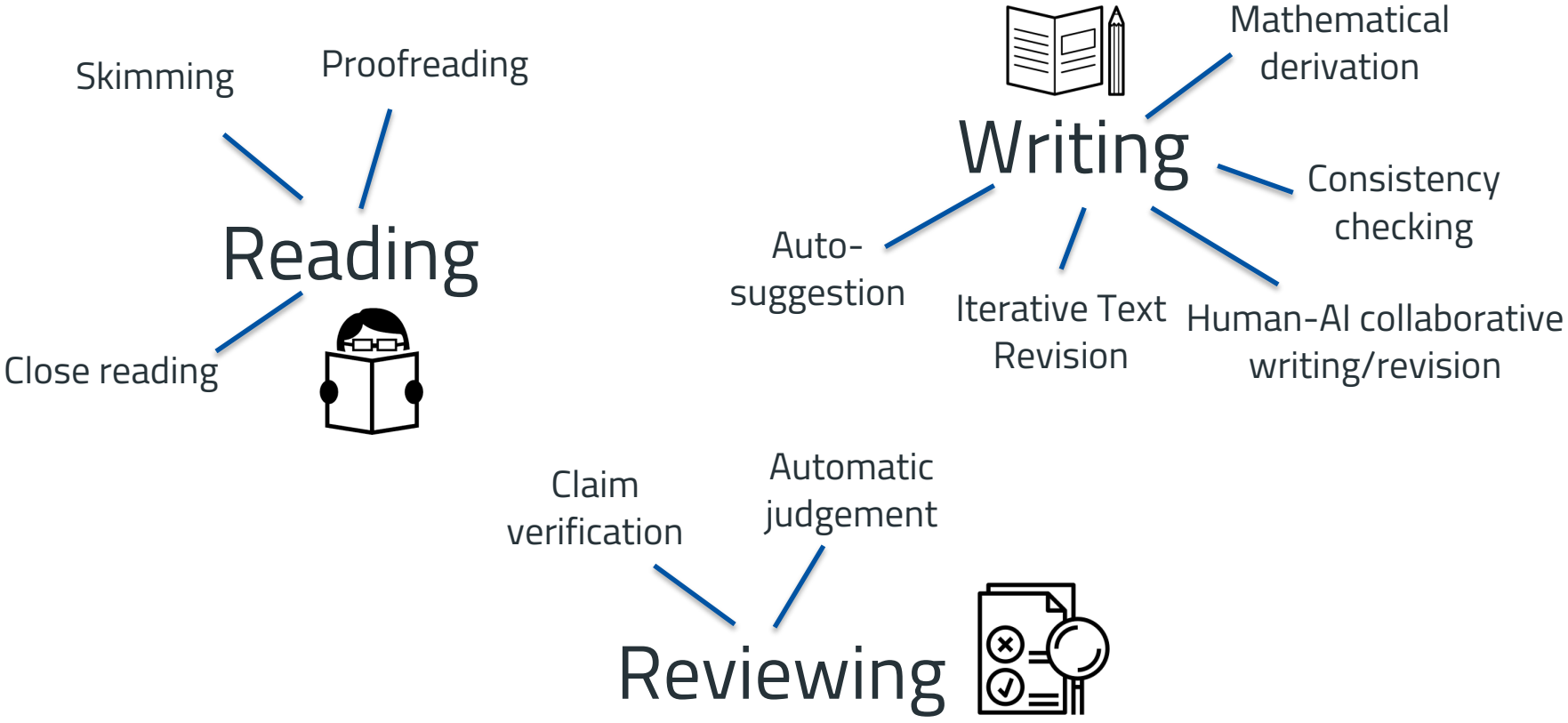
Pushing toward the expert-level AI

- Understand experts' writing and thinking process at workplaces
- Develop and design interactive systems to facilitate collaboration between human experts (e.g., scientists, lawyers) and AI tools.
- Create complex, compositional, and domain-specific expert-level benchmarks

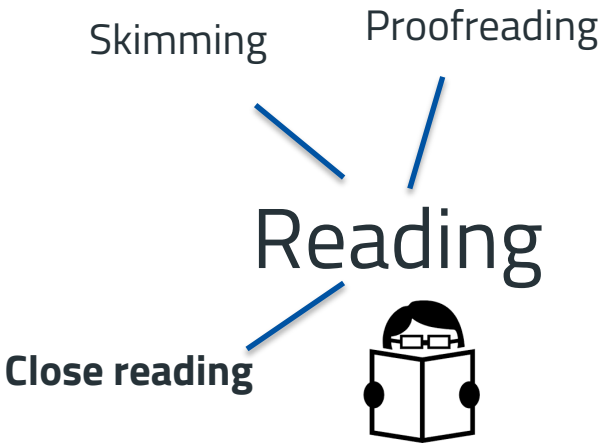


Forecasting the future of artificial intelligence with machine learning-based link prediction in an exponentially growing knowledge network, 2023

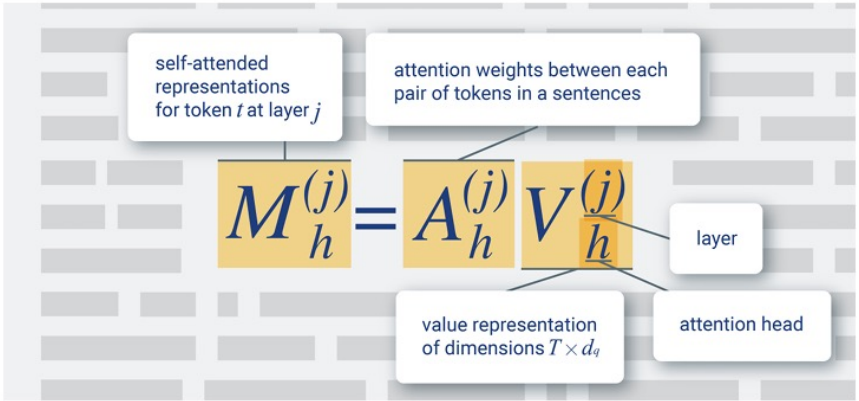
Improving scientific research with interactive NLP systems



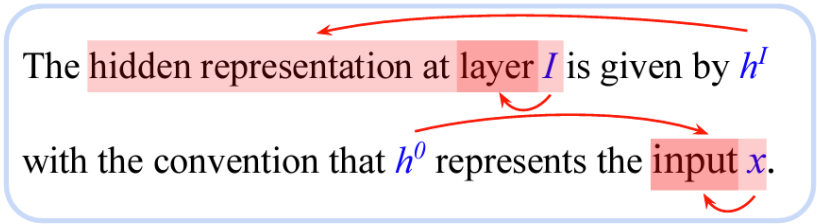
Improving scientific reading with interactive NLP systems



- Augmented PDF reader with interactive interfaces
- Provide in-context definitions of terms & symbols.



ScholarPhi (CHI'21); Semantic Reader (ACMC'23)



HEDDEx (SDP@EMNLP'20); TaDDEx (under review)

Glossary of key terms

Listed in order of appearance.

SRL: semantic role labeling.

LISA: linguistically-informed self-attention.

D&M: parses predicted by Dozat and Manning (2017), the winner of the 2017 CoNLL shared task.

WSJ: Wall Street Journal.

SA: a version of our own self-attention model which does not incorporate syntactic information.

CoNLL-2005: dataset (Carreras and Màrquez, 2005) based on the original PropBank corpus (Palmer et al., 2005), which labels the Wall Street Journal portion of the Penn TreeBank corpus (PTB) (Marcus et al., 1993) with predicate-argument structures, plus a challenging out-of-domain test set derived from the Brown corpus (Francis and Kučera, 1964).

+Gold: gold syntactic parses.

POS: part-of-speech.

UAS: unlabeled attachment scores.

LAS: labeled attachment scores.

L: LISA.

D: D&M.

+: parses were completely correct.

PP: prepositional phrase.

E: ELMo embeddings.

+D&M: parses predicted by Dozat and Manning (2017), the winner of the 2017 CoNLL shared task.

-: parses were completely incorrect.

PTB: Penn TreeBank.

SGD: stochastic gradient descent.

linguistically-informed self-attention: a variation of self-attention that combines multi-head self-attention with multi-task learning across dependency parsing, part-of-speech tagging, predicate detection and SRL.

Fix Labels: a correction to model predictions that fixes labels on spans matching gold boundaries.

Merge Spans: a correction to model predictions that merges adjacent predicted spans into a gold span.

Split Spans: type of span boundary error.

Fix Span Boundary: type of span boundary error.

MTL: multi-task learning.

PoE: ensemble model from He et al. (2017).

Glossary of key symbols

Listed in order of appearance.

r: input for a joint predicate/POS classifier.

f: index; frame.

s_{ft}: role label scores for the token at index t with respect to the predicate at index f; unary scores.

y_t^{dep}: dependency labels.

λ_1: penalty on the syntactic attention loss.

V_{parse}: token values; value representation; value representations.

$T^{(j)}(\cdot)$: jth attention layer.

$s_t^{(j)}$: output of layer.

H: number of attention heads; number of self-attentions.

h: attention head.

5. Glossary

Deep Speech 2: End-to-End Speech Recognition in English and Mandarin

Baidu Research – Silicon Valley AI Lab*
Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Erich Elsen, Jesse Engel, Linxi Fan, Christopher Fougner, Tony Han, Awni Hannun, Billy Jun, Patrick LeGresley, Libby Lin, Sharan Narang, Andrew Ng, Sherjil Ozair, Ryan Prenger, Jonathan Raiman, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Yi Wang, Zhiqian Wang, Chong Wang, Bo Xiao, Dani Yogatama, Jun Zhan, Zhenyao Zhu

Abstract

We show that an end-to-end deep learning approach can be used to recognize either English or Mandarin Chinese speech—two vastly different languages. Because it replaces entire pipelines of hand-engineered components with neural networks, end-to-end learning allows us to handle a diverse variety of speech including noisy environments, accents and different languages. Key to our approach is our application of HPC techniques, resulting in a 7x speedup over our previous system [26]. Because of this efficiency, experiments that previously took weeks now ru… chitect… with th… Finally… show t… ing low…

1 Introduc…

Decades worth… matic speech re… ASR models en… present the seco… to-end learning… Mechanical Tur… modification, ar… a single ASR sy… mans. Since ou… learning techni… computing, and… through these te… English by up t…

One of the chall…

[Kang et al., SDP 2020 arxiv./2010.05129; Martin et al., arxiv/2305.14660]

Coordination Resolution in Definition Extraction

And the top-left corner and the bottom-right corner of the predicted projected box are $(i - S\hat{o}_{t_{i,j}}, j - S\hat{o}_{l_{i,j}})$ *and* $(i + S\hat{o}_{b_{i,j}}, j + S\hat{o}_{r_{i,j}}])$ *respectively.*

[Kang et al., SDP 2020 arxiv./2010.05129; Martin et al., arxiv/2305.14660]

Coordination Resolution in Definition Extraction

And the top-left corner and the bottom-right corner of the predicted projected box are $(i - S\hat{o}_{t_{i,j}}, j - S\hat{o}_{l_{i,j}})$ *and* $(i + S\hat{o}_{b_{i,j}}, j + S\hat{o}_{r_{i,j}}])$ *respectively.*

Top-left corner of the predicted projected box

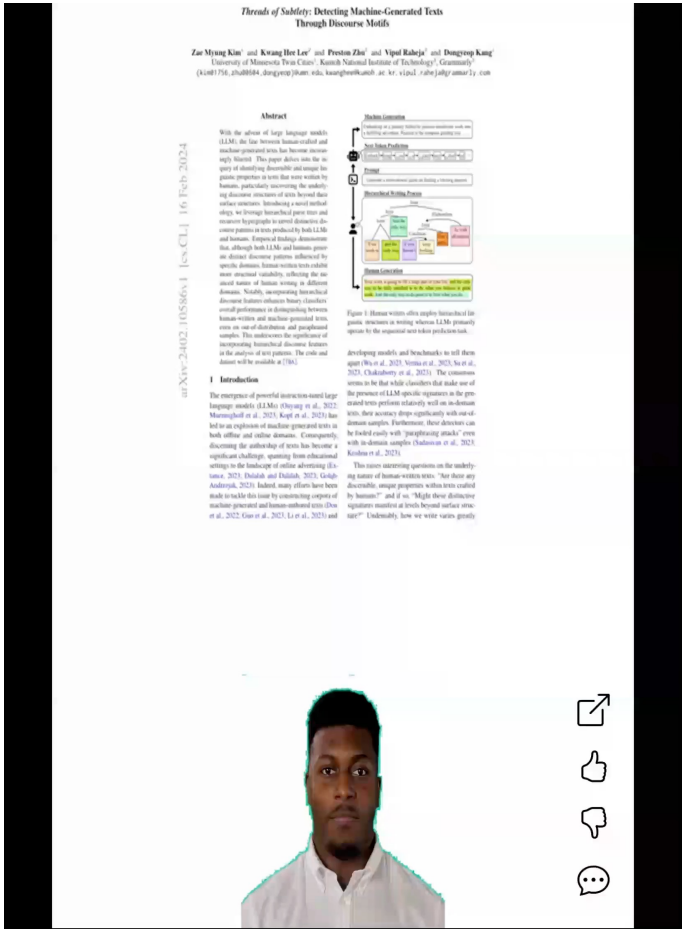
[Kang et al., SDP 2020 arxiv./2010.05129; Martin et al., arxiv/2305.14660]

Coordination Resolution in Definition Extraction

And the top-left corner and the bottom-right corner of the predicted projected box are $(i - S\hat{o}_{t_{i,j}}, j - S\hat{o}_{l_{i,j}})$ *and* $(i + S\hat{o}_{b_{i,j}}, j + S\hat{o}_{r_{i,j}}])$ *respectively.*

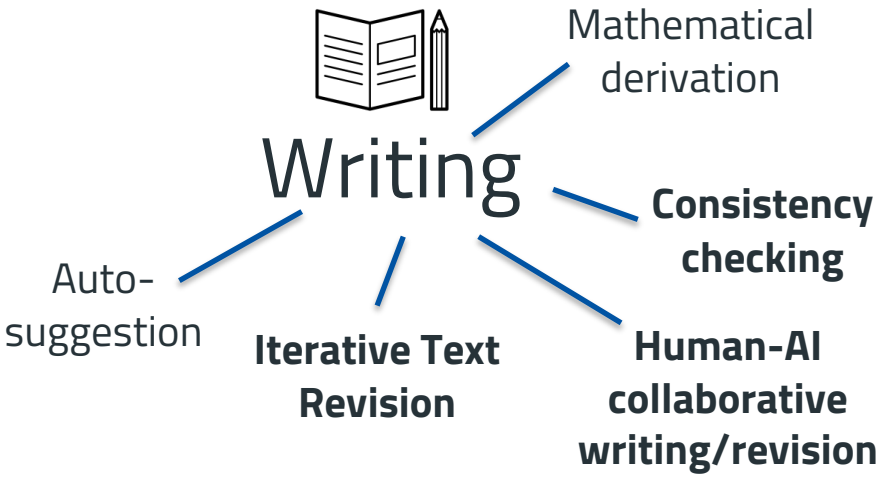
Bottom-left corner of the predicted projected box

SciTok: A Short-form Paper "Watching" for Easy Science



Improving scientific writing with interactive NLP systems

CoEdit (EMNLP Findings 23)

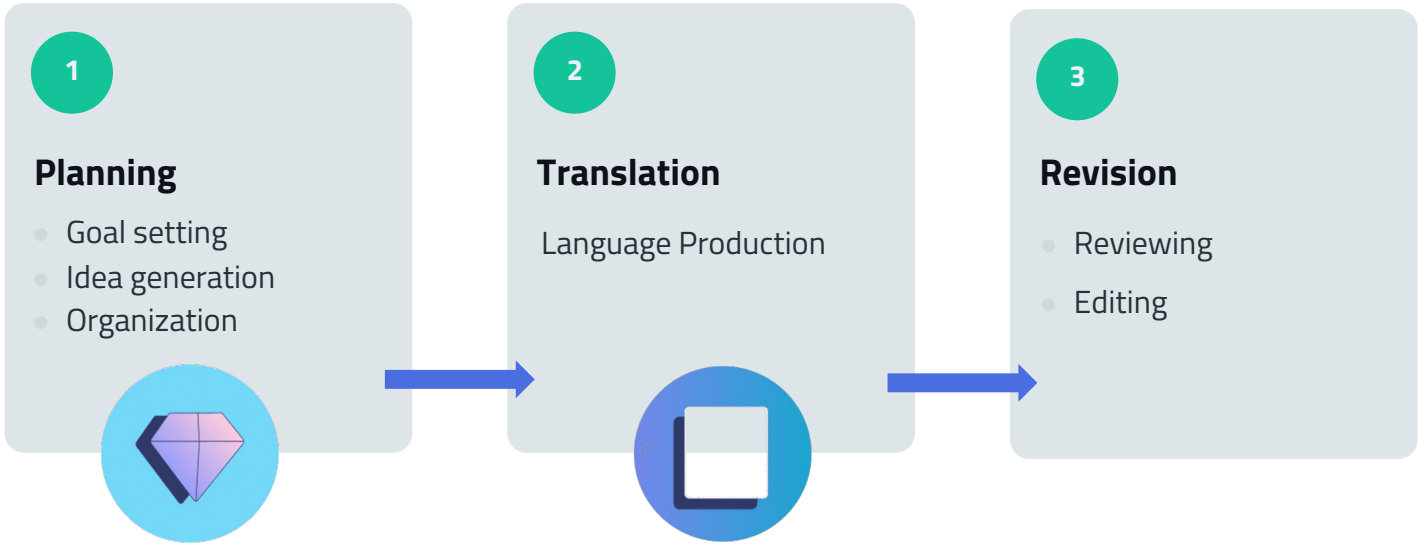



```
disambiguation \cite{banerjee2002adapted,huang-etal-20
Prior work in automated definition detection has addre
\cite{reiplinger-etal-2012-extracting,jin-etal-2013-mi
tically,vanetik-etal-2020-automated,Veyseh2020AJM}.
Definition extraction is especially important for scho
unfamiliar technical terms that readers must understan
```

Consistency and coherence checker

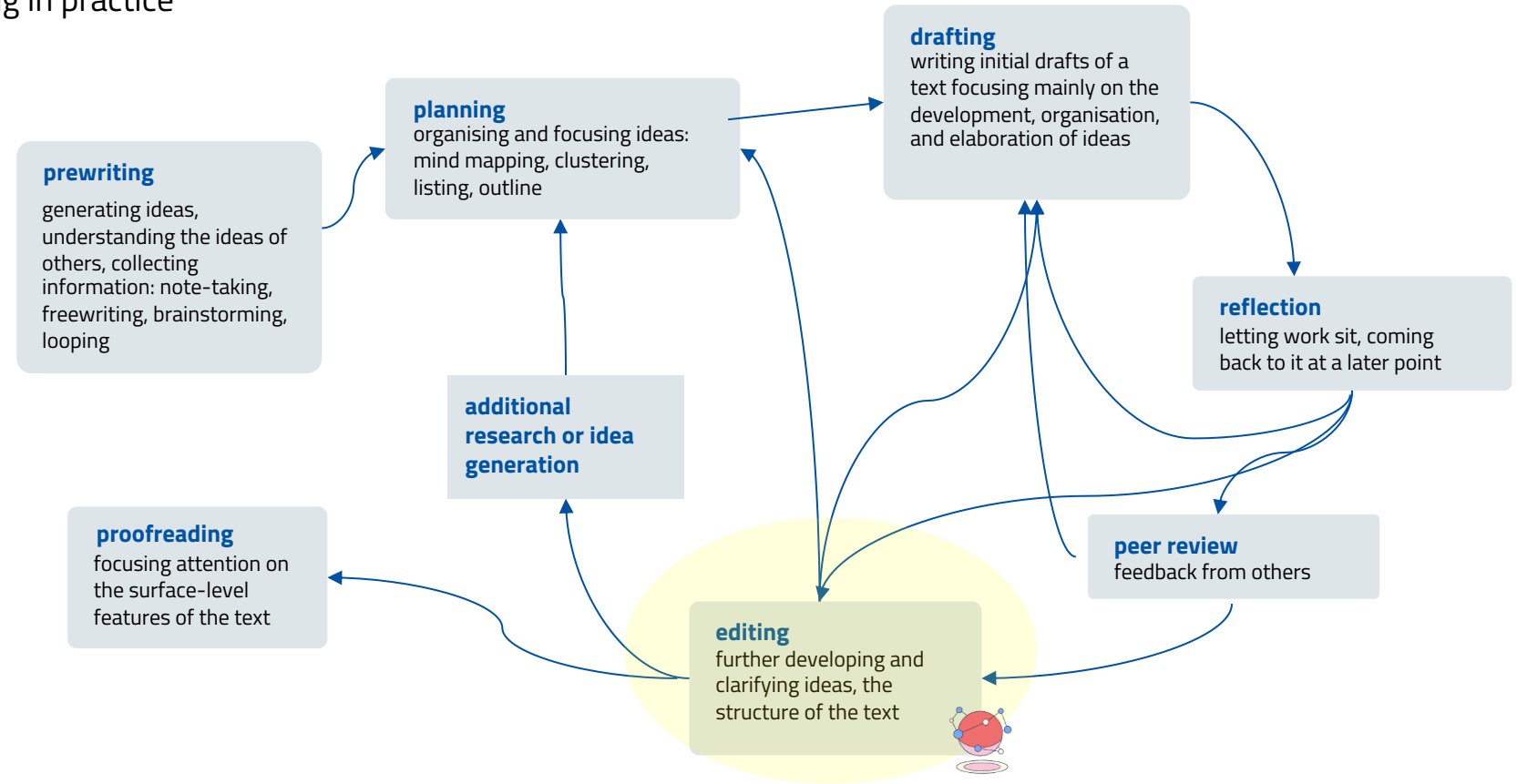
- Collaborative text editing via Instruction Tuning (CoEdit)
- Checking consistency and coherence of your writing

Writing in theory



Flower, Linda, and Hayes, John R.. "A Cognitive Process Theory of Writing." College Composition and Communication, 1981

Writing in practice



[Du et al., ACL 2022 arxiv/2203.03802]

Learning from *iterative human writing*

v1

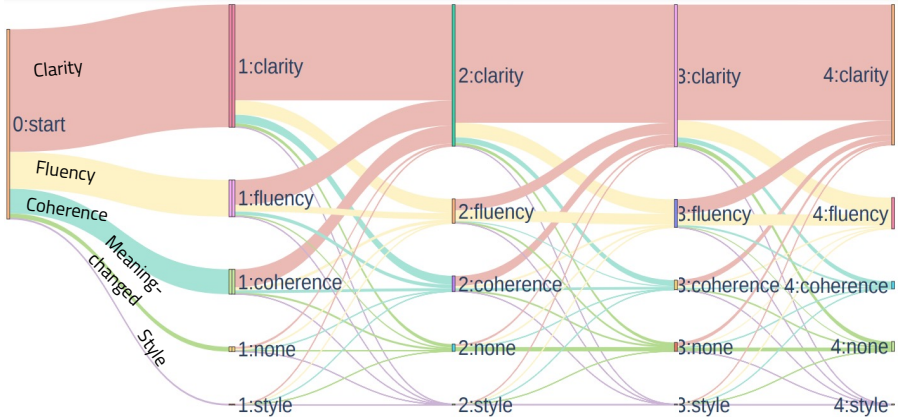
Each comment was annotated by three different annotators, which achieved high inter-annotator agreement. The proposed annotation {~~process~~ approach} CLARITY is also language and domain independent{, nevertheless, it was currently applied for Brazilian Portuguese} MEANING-CHANGED .

v2

Each comment was annotated by three different annotators, {~~which~~ and} COHERENCE achieved high inter-annotator agreement. The {new} MEANING-CHANGED proposed annotation approach is also language and {~~domain independent, nevertheless, it was currently~~domain-independent (although it has been} CLARITY applied for Brazilian Portuguese{)} FLUENCY .

v3

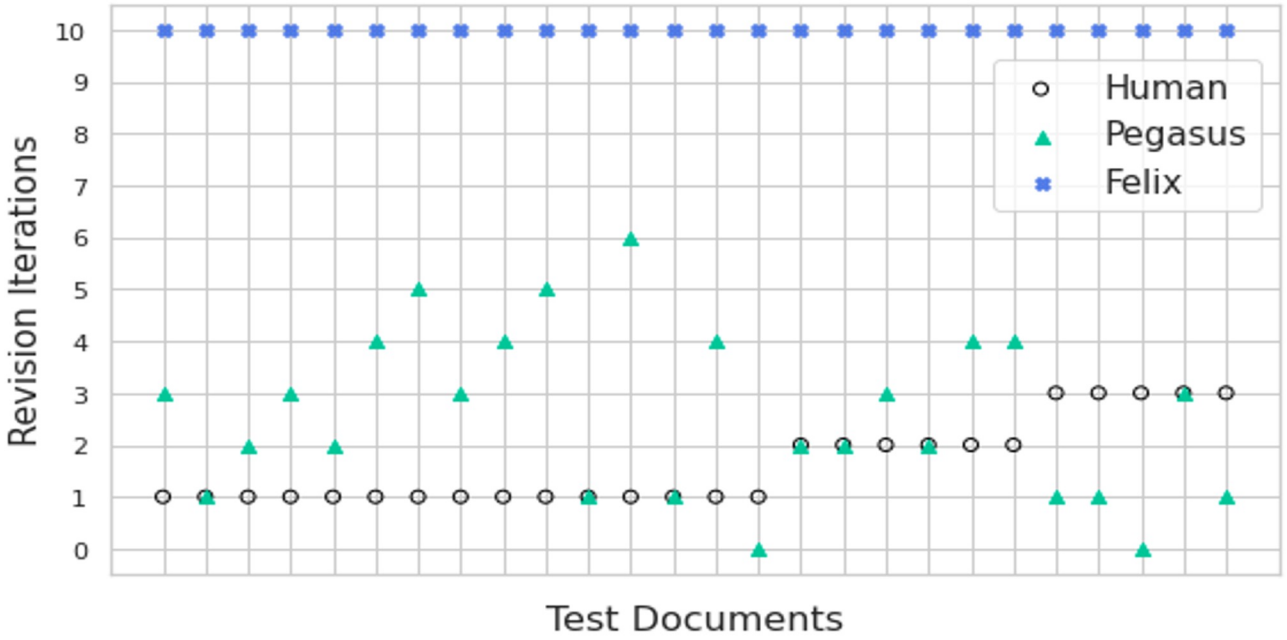
Each comment was annotated by three different annotators {~~,~~} FLUENCY and achieved high inter-annotator agreement. The {~~new~~} COHERENCE proposed annotation approach is also language and domain-independent {~~(although it has been applied~~ nevertheless it is currently customized} COHERENCE for Brazilian Portuguese {~~)~~} FLUENCY .



Trajectory of edit intentions in Iterative Writing (ACL'22; EMNLP'22)

[Du et al., ACL 2022 arxiv/2203.03802]

Can models learn when to stop iterating?



Writing in practice

prewriting
generating ideas, understanding the ideas of others, collecting information: note-taking, freewriting, brainstorming, looping

planning
organising and focusing ideas: mind mapping, clustering, listing, outline

drafting
writing initial drafts of a text focusing mainly on the development, organisation, and elaboration of ideas

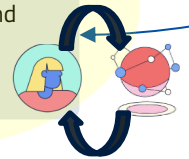
reflection
letting work sit, coming back to it at a later point

peer review
feedback from others

editing
further developing and clarifying ideas, the structure of the text

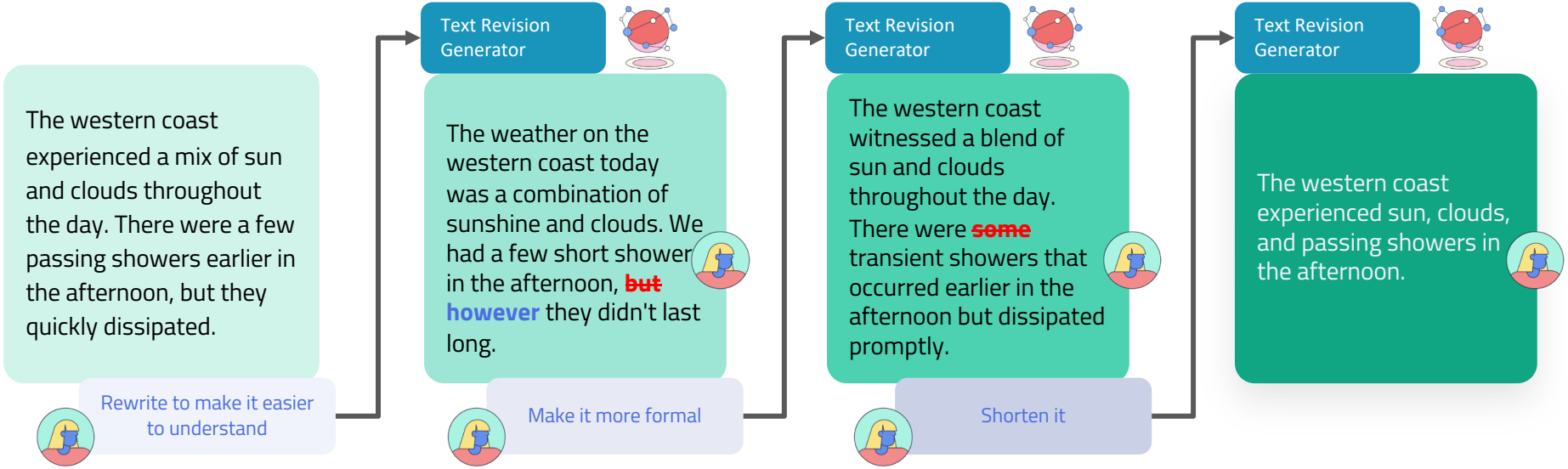
additional research or idea generation

proofreading
focusing attention on the surface-level features of the text



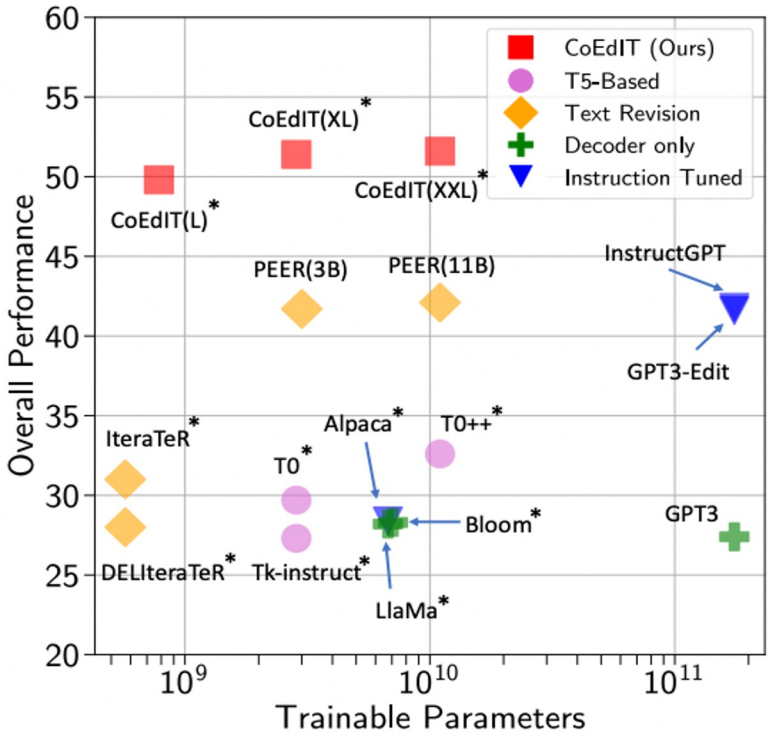
[Raheja et al., EMNLP Findings 23 arxiv/2305.09857]

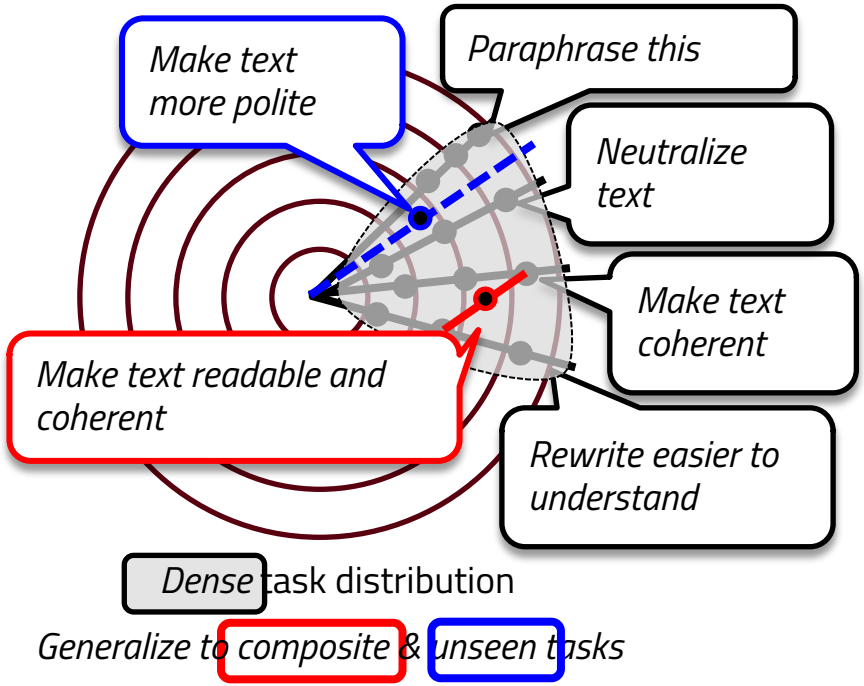
Conversational Text Editing via Instruction Tuning



[Raheja et al., EMNLP Findings 23 arxiv/2305.09857]

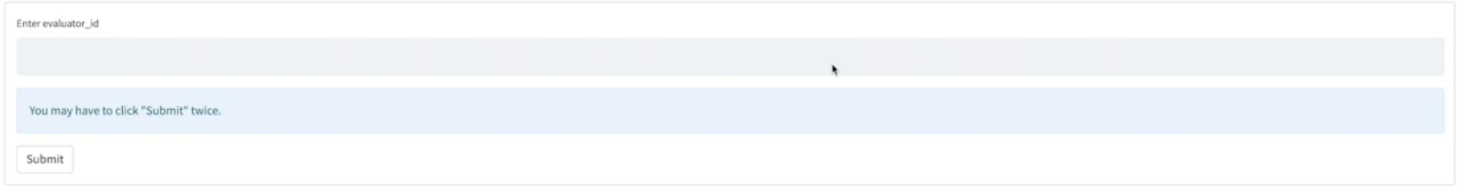
Conversational Text Editing via Instruction Tuning





[Du&Kim et al., in2writing 2022 arxiv/2204.03685]

Human-AI collaborative text revision



Writing in practice

prewriting
generating ideas, understanding the ideas of others, collecting information: note-taking, freewriting, brainstorming, looping

planning
organising and focusing ideas: mind mapping, clustering, listing, outline

drafting
writing initial drafts of a text focusing mainly on the development, organisation, and elaboration of ideas

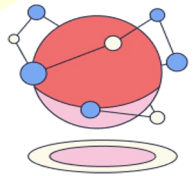
reflection
letting work sit, coming back to it at a later point

peer review
feedback from others

editing
further developing and clarifying ideas, the structure of the text

proofreading
focusing attention on the surface-level features of the text

additional research or idea generation



[Koo&Martin et al., in2writing 2023 arxiv/ 2304.00121]

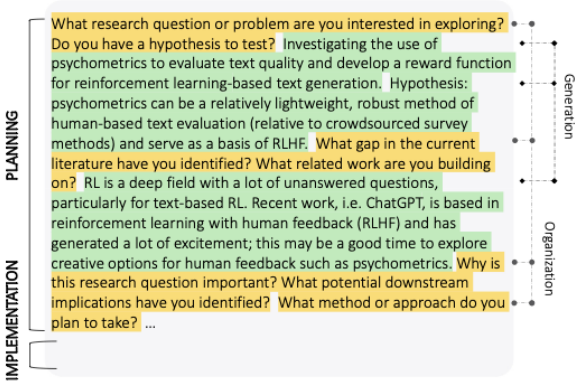
A Dataset of Writing Trajectory

Feb 02 2023 11:39:00:562

```
\documentclass{article}
\usepackage[utf8]{inputenc}

\title{Research Goals}
\date{February 2023}

\begin{document}

\maketitle

\section{Outline}
\begin{enumerate}
  \item Individualized Prediction
  \item
\end{enumerate}


\end{document}
```

Label	Description
PLANNING The writer's intention is to get their ideas down on paper in a semi-structured manner.	
generation	The process of adding ideas to the document.
organization	Structuring the generated concepts.
IMPLEMENTATION The writer's intention is to produce high-quality and persuasive text that meets their writing goals.	
lexical chaining	Writing coherent text where sentences are linked by the semantic relationships between words [8].
REVISION The writer's intention is to improve the clarity, consistency, coherence, and style of the written text.	
syntactic	Fixing grammar, spelling, and punctuation.
lexical	Changing words to clarify meaning or improve coherence.
structural	Reordering text to improve organization.

Table 1: Simplified annotation schema applied to our dataset



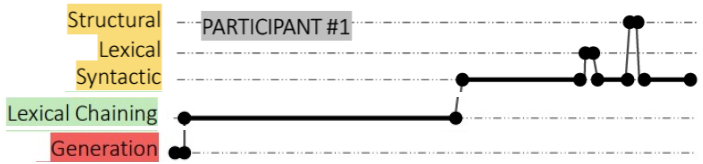
Capture all stages of the writing process in a short, prompt-given essay (< 30 min)

→

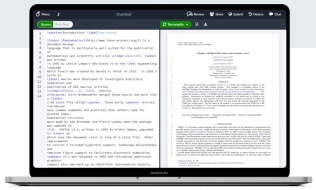
Taxonomize and **Annotate** writing patterns in academic writing

↓

Transform it into information that can aid writing assistants to provide better feedback



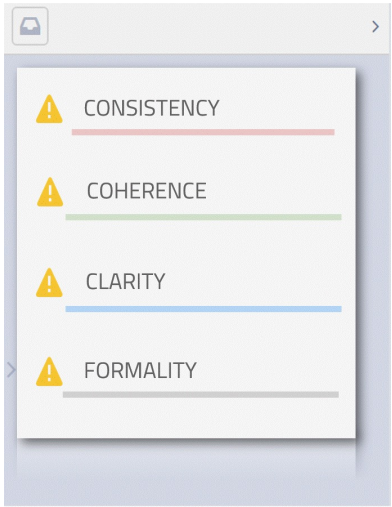
Planning → Implementation → Revision



Intelligent Writing Assistant for Scientific Writers

```
disambiguation \cite{banerjee2002adapted,huang-etal-20
Prior work in automated definition detection has addre
\cite{reiplinger-etal-2012-extracting,jin-etal-2013-mi
tically,vanetik-etal-2020-automated,Veyseh2020AJM}.
Definition extraction is especially important for scho
unfamiliar technical terms that readers must understan

```

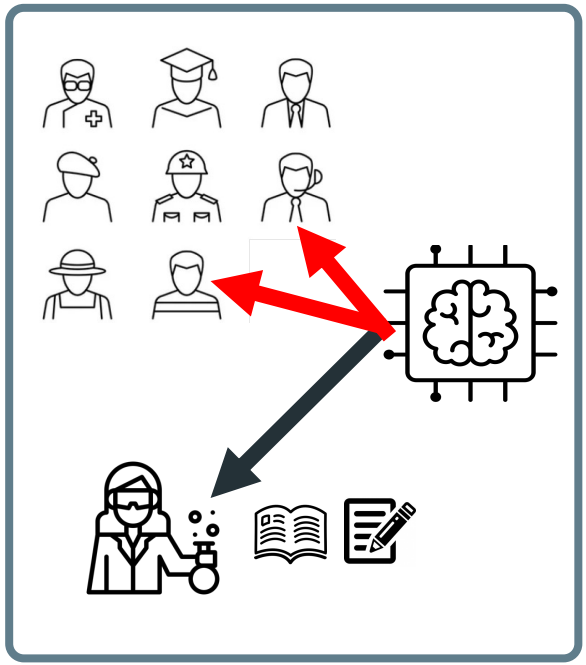
Consistency checker



Automatic readibility scoring

```
Particularly, the symbol
```

Discourse-aware auto-suggestion



Pushing toward the expert-level AI

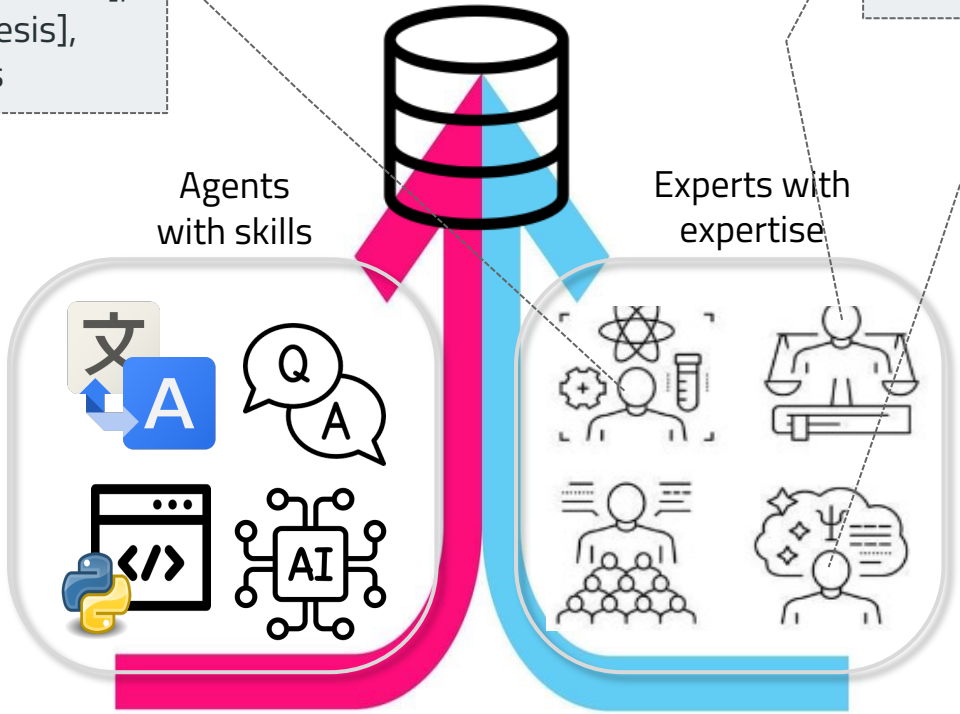
- Understand experts' writing and thinking process at workplaces
- Develop and design interactive systems to facilitate collaboration between human experts (e.g., scientists, lawyers) and AI tools.
- Create complex, compositional, and domain-specific expert-level benchmarks

Create ExpertBench by collecting skillsets of professions from domain experts

Legal experts need [search], [summarization], [knowledge synthesis], [negotiation] skills

Teachers need [communication], [presentation], [teaching] skills

Researchers need [critical thinking], [research], [contrast], [creative thinking] skills

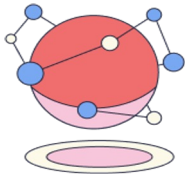


(1) Collecting c*omplex, chained tasks* by multi-agent collaboration with minimal expert intervention (bottom-up)

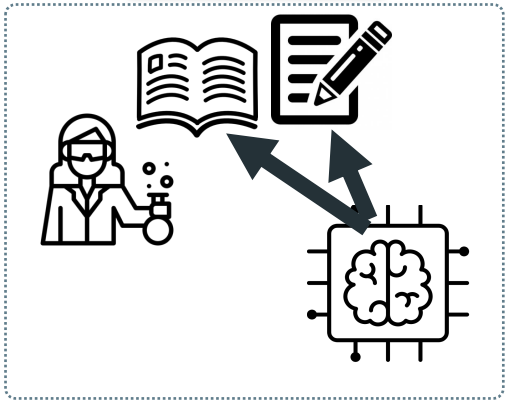
(2) Collecting *professional knowledge and skills* from human experts (top-down)

Takeaways

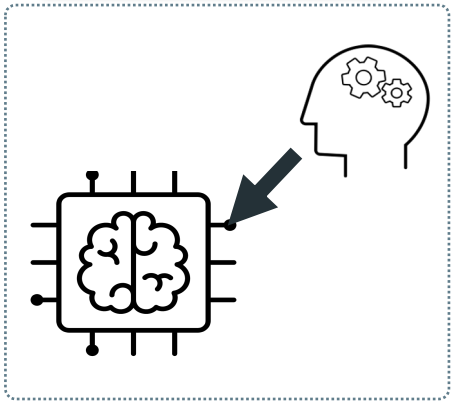
- Modeling each process of human writing (iterative editing, planning, augmentation) is **extremely challenging**.
- **Human-AI collaborative writing** improves the control of interaction, revision quality, and evaluation to either party.
- Any intelligent interface runs the risk of creating a false sense of clarity.
  - Thought-terminating (Thi Nguyen (2021)) by AI suggestions is detrimental to science.
- Collect **high-quality expert benchmarks** and develop carefully designed interfaces for supporting thinking and writing processes by (knowledge-based) domain experts



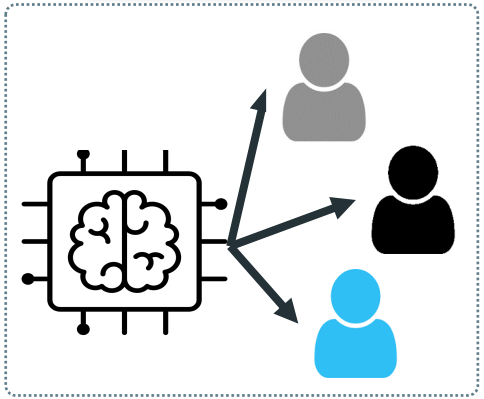
Expert-level AI

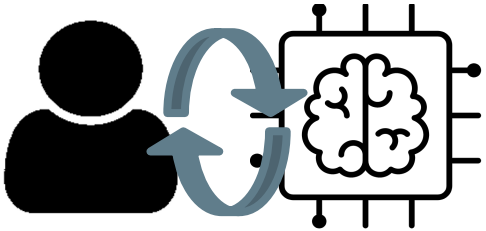


Cognitive Scaffolding

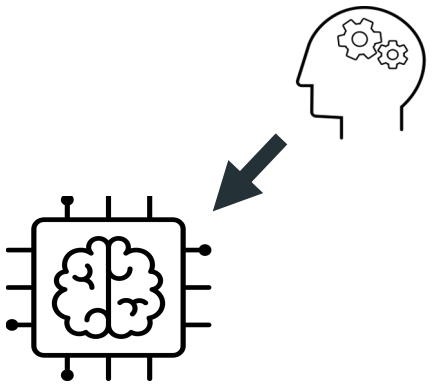


Societal Alignment





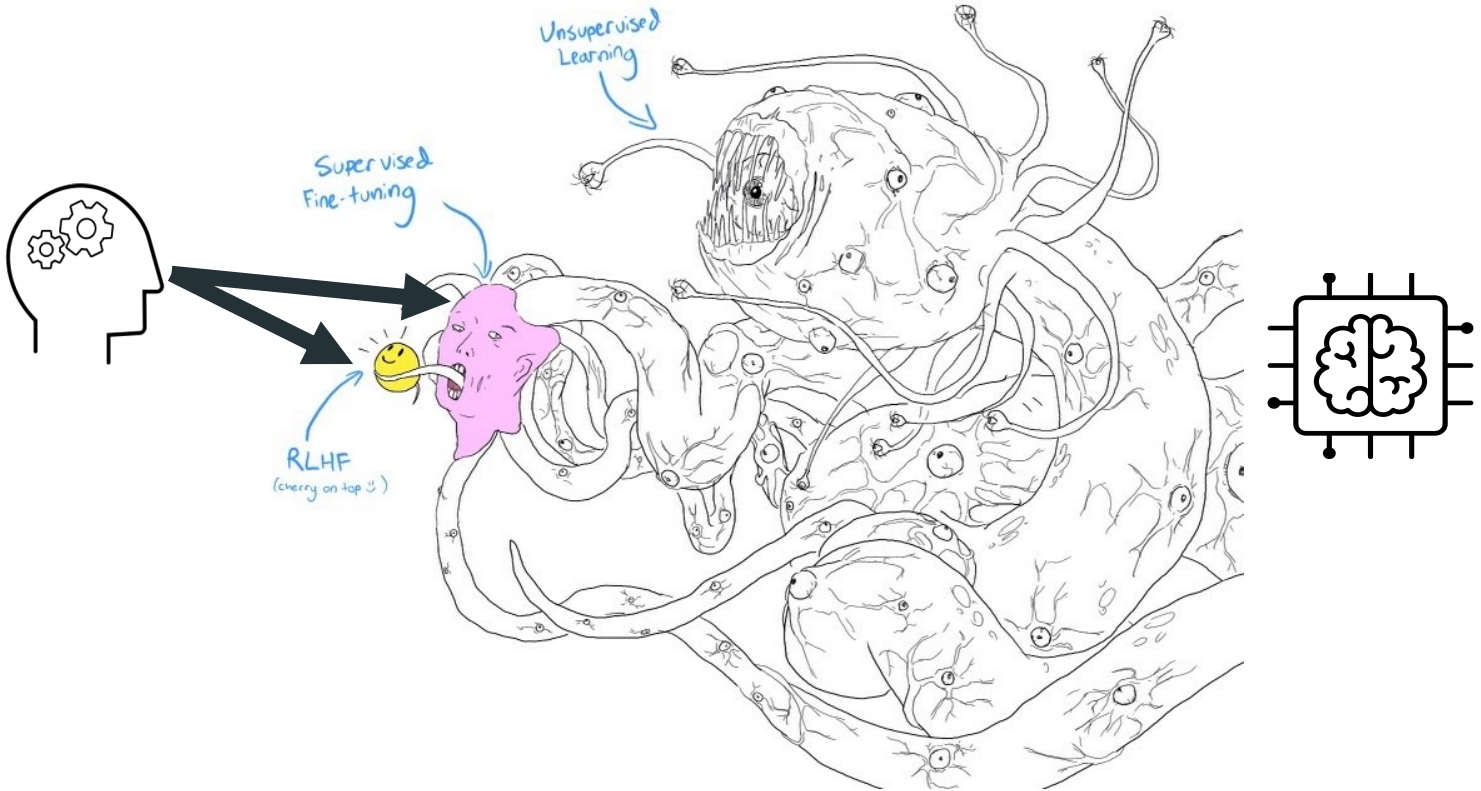
Human-centric NLP



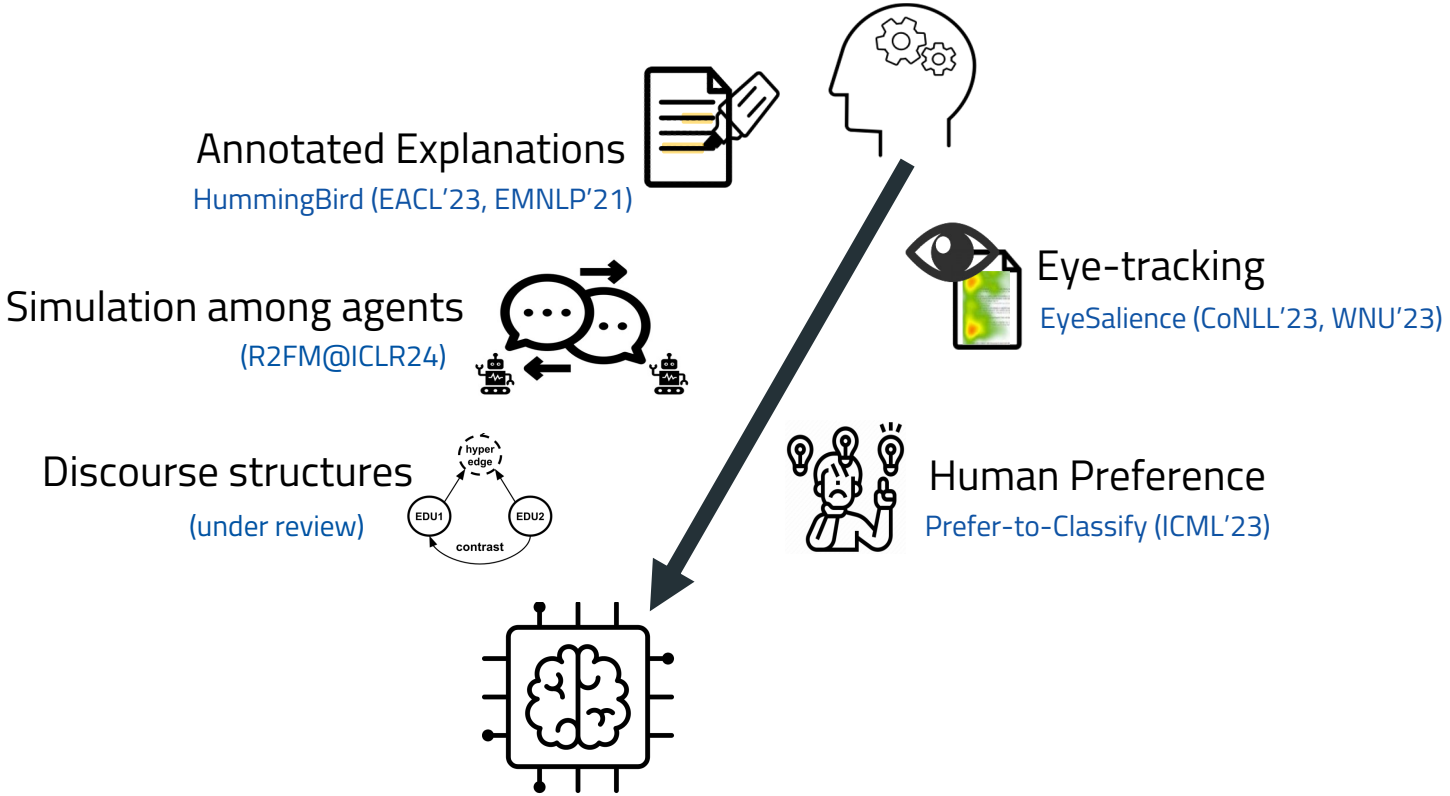
Cognitively Scaffolding LLMs

- Collect different types of human cognition signals and develop cognitively-inspired AI models
- Support thinking process in advanced writing tasks (persuasive framing) and improve thematic coherence and controllability in long-form compositional writing tasks (storytelling)

Scaffold LLM with human cognition for better human alignment

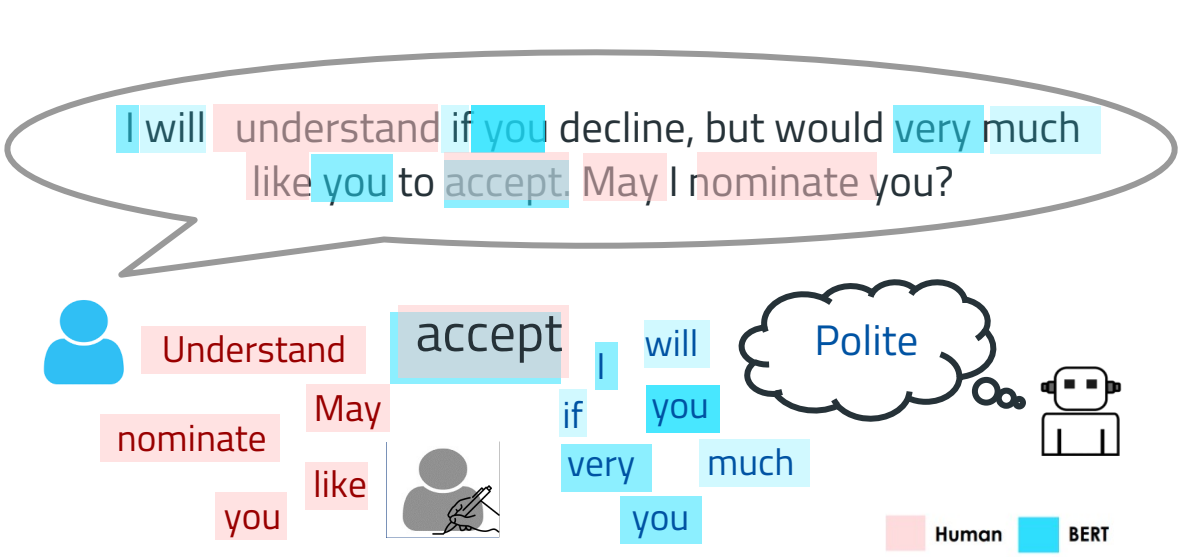


Scaffold LLM with human cognition for better human alignment

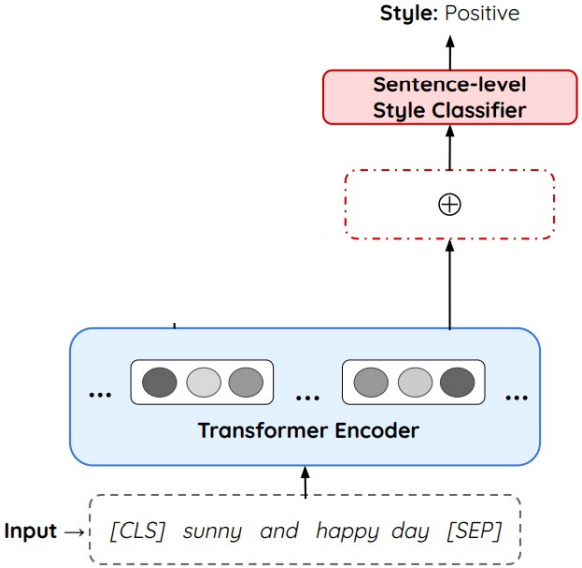


[Hayati et al., EMNLP'21 arxiv/2109.02738; Hayati et al., EACL'23 arxiv/2210.07469]

Learning human cognition from annotated lexical explanations



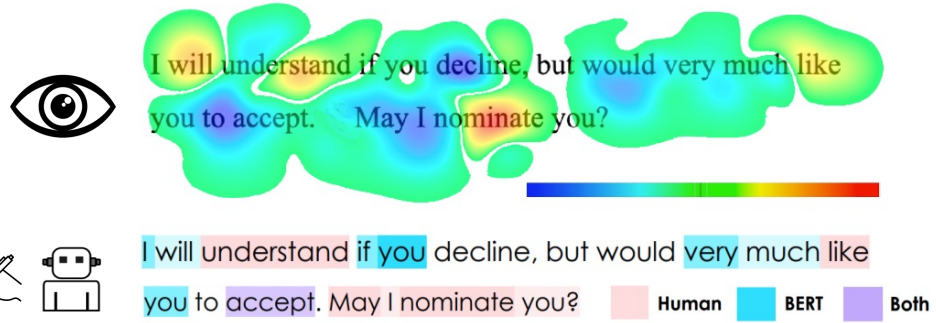
Does BERT Learn as Humans Perceive?



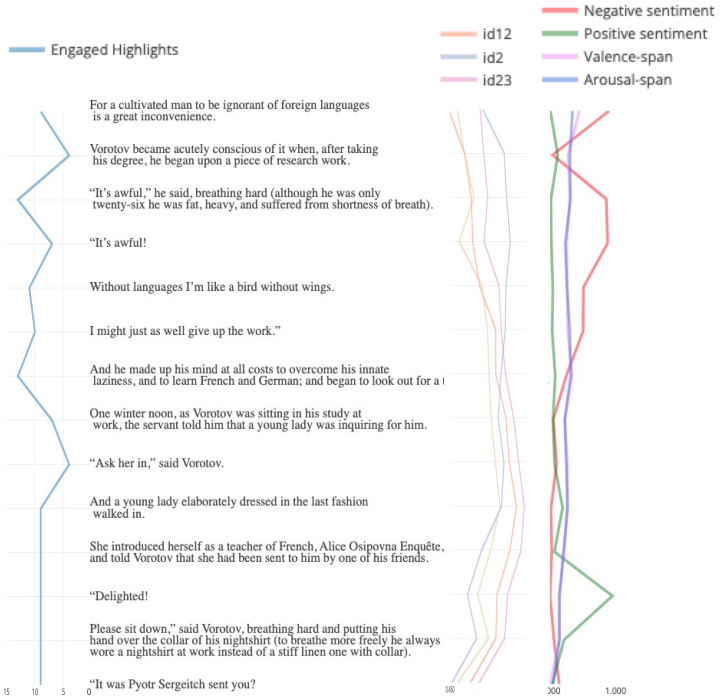
Interpretable AI System

[de Langis et al., CoNLL'23 arxiv/2212.09873v1; Neis et al., WNU @ACL23]

Learning human cognition from eye movement



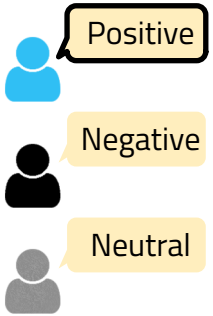
Eye-tracking for Textual Saliency



Reader Engagement in Fiction Literacy

[Kim et al., ICML 23 arxiv/2306.04925]

Learning human cognition from preference feedback



A: I got 3 veggies and a side of fries for over a 11 dollars if you like homecooked food

B: She listened to my ideas, asked questions to get a better idea about my style, and was excellent at offering advice as if I were a total pleb.

Annotation Records

Extractive

Subjective

Crowd Workers

LLMs

Generative

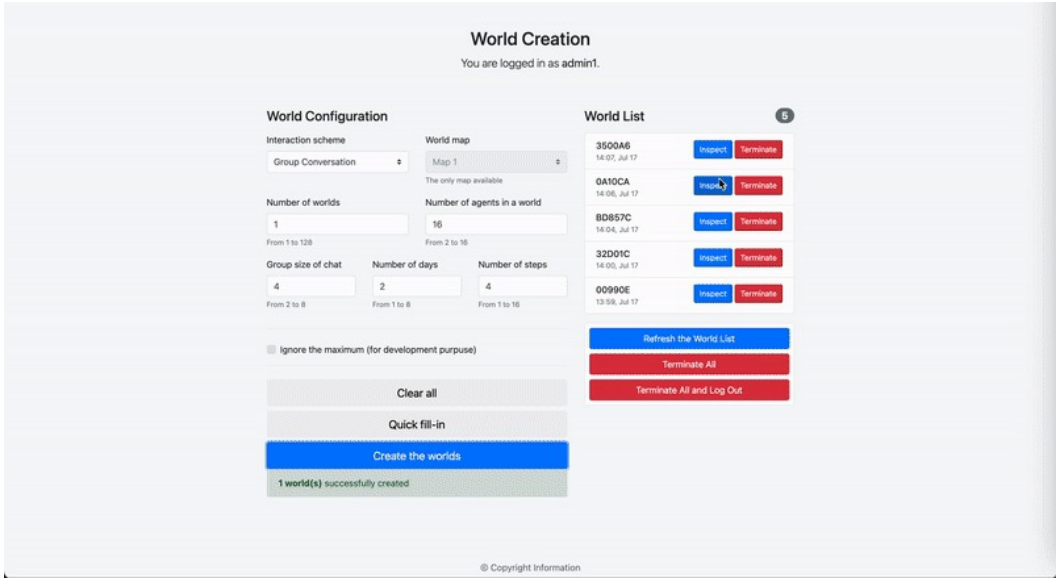
A is preferably more positive than B. (A > B)

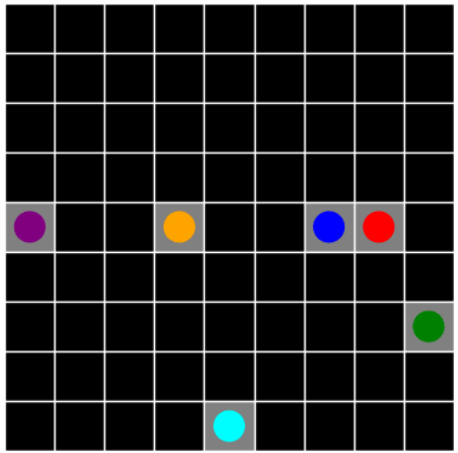
B is preferably more positive than A. (A < B)

B is preferably more positive than A. (A < B)

[Yang et al., R2FM @ICLR 24]

Learning human cognition from simulated interaction among agents





agent 1: I am agent 1. Given that agent 2 is at x=7, y=3, agent 3 is at x=2, y=1, and agent 4 is at x=7, y=4, I will continue searching the upper-left region. So I will go UP.

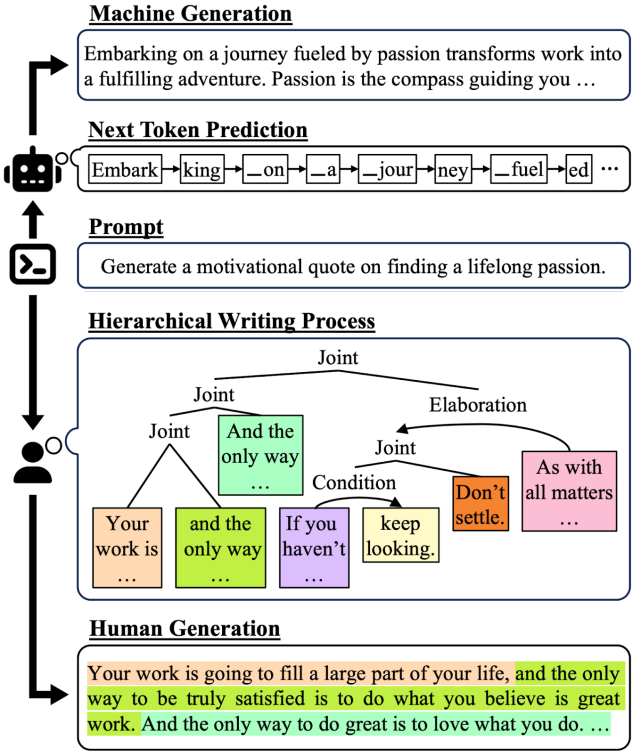
referee: Agent 1 did not move and stayed at x=3, y=7 because the destination was out of range.

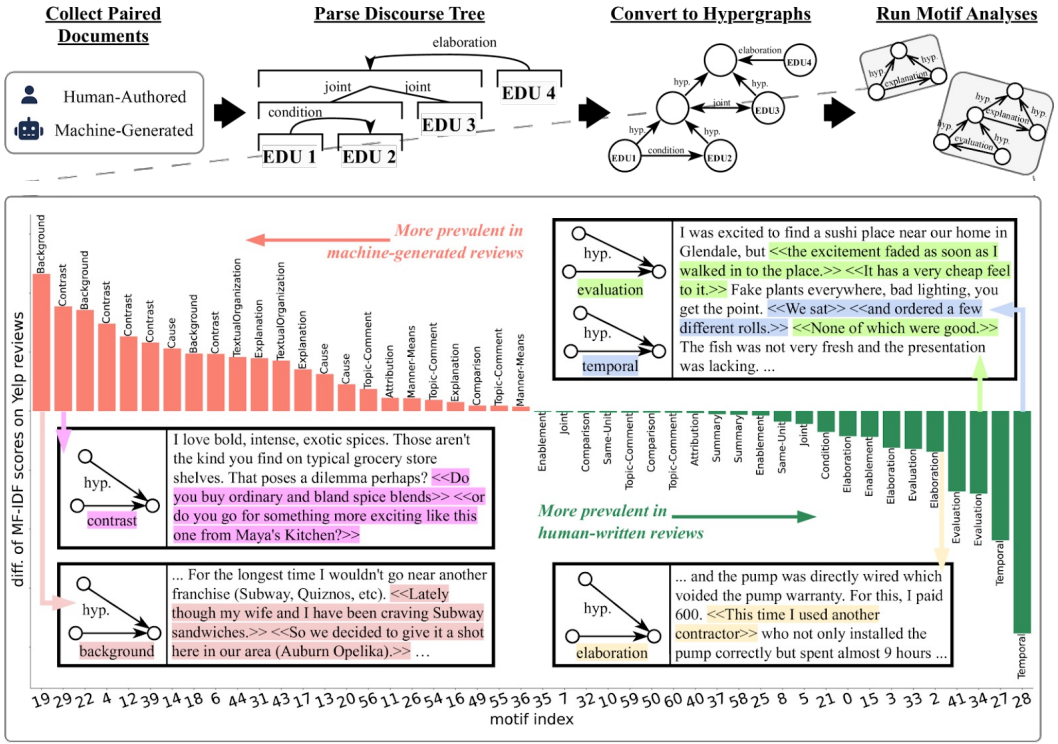
agent 2: I am agent 2. Given that agent 1 is searching the upper-left region and staying at x=3, y=7, I will update my strategy. Since agent 3 is searching the upper-right region and staying at x=2, y=1, and agent 4 is searching the lower-right

SocialSyms: Multi-agent Simulation Platform for Social Reasoning and Collaborative Tasks

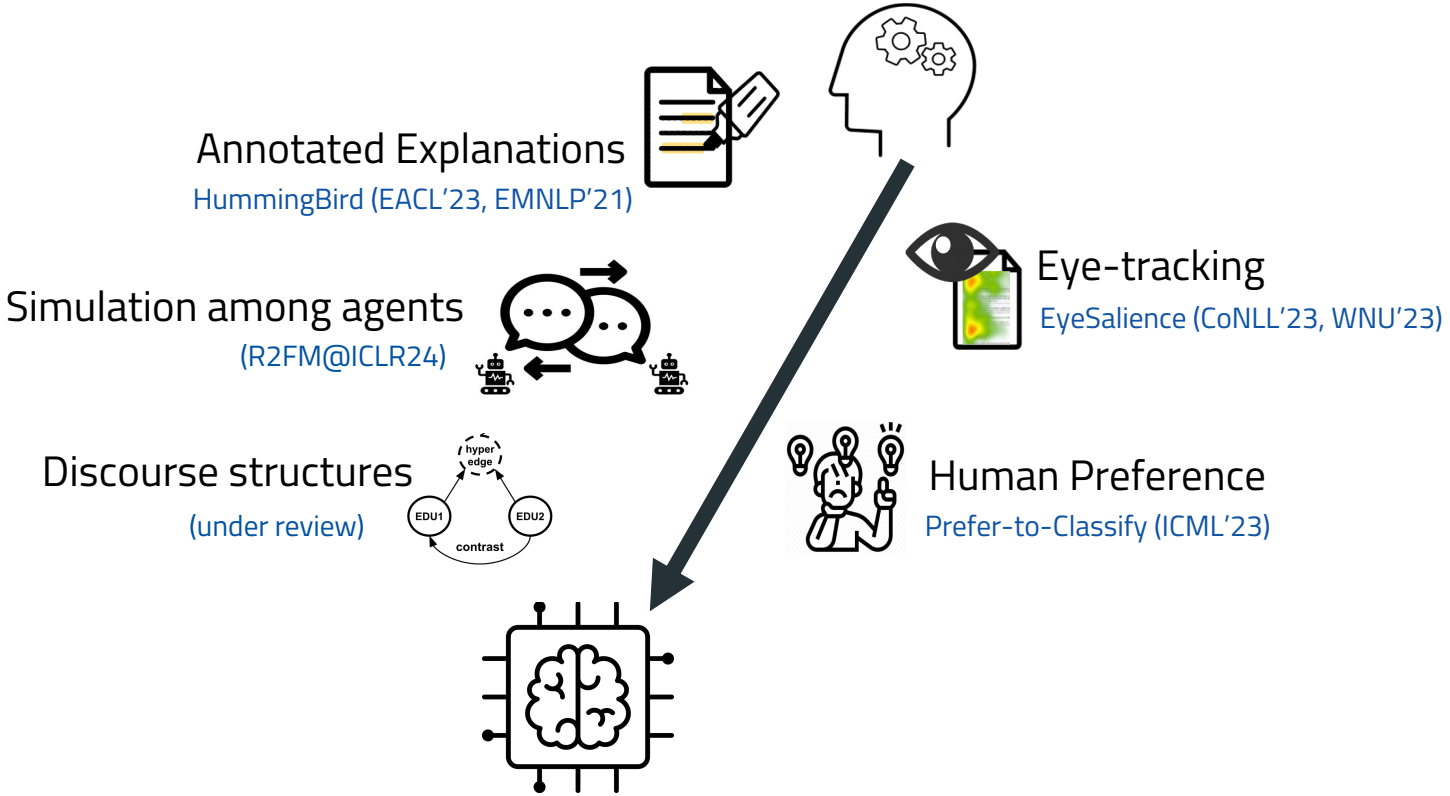
[Kim et al., under review arxiv/2402.12255]

Learning human cognition from discourse structures of human writing

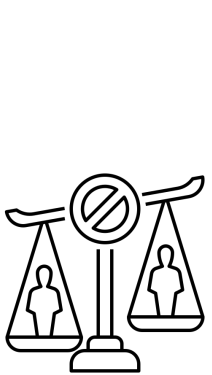




Scaffold LLM with human cognition for better human alignment



Scaffold LLM with ~~human cognition~~ **machine-generated data** causes to produce biases and artifacts

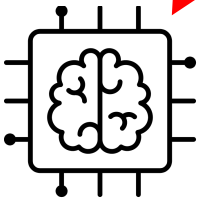


Biases in LLM-based evaluation

CoBBLeR (under review)

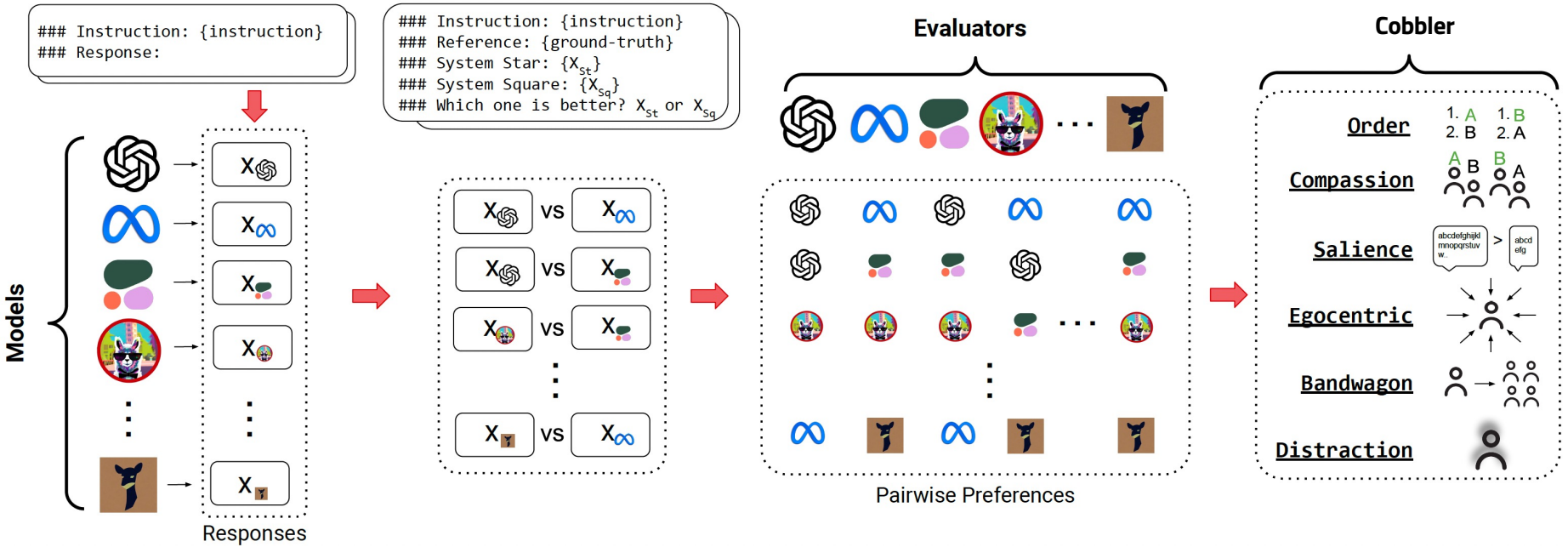
Tracking Artifactuality in AI ecosystem

Under Surface (under review)



[Koo et al., under review arxiv/2309.17012]

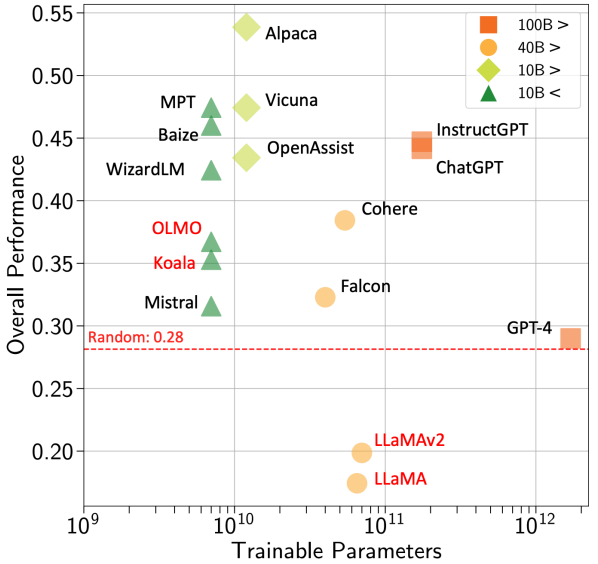
Analyzing cognitive biases of LLM-based evaluation



[Koo et al., under review arxiv/2309.17012]

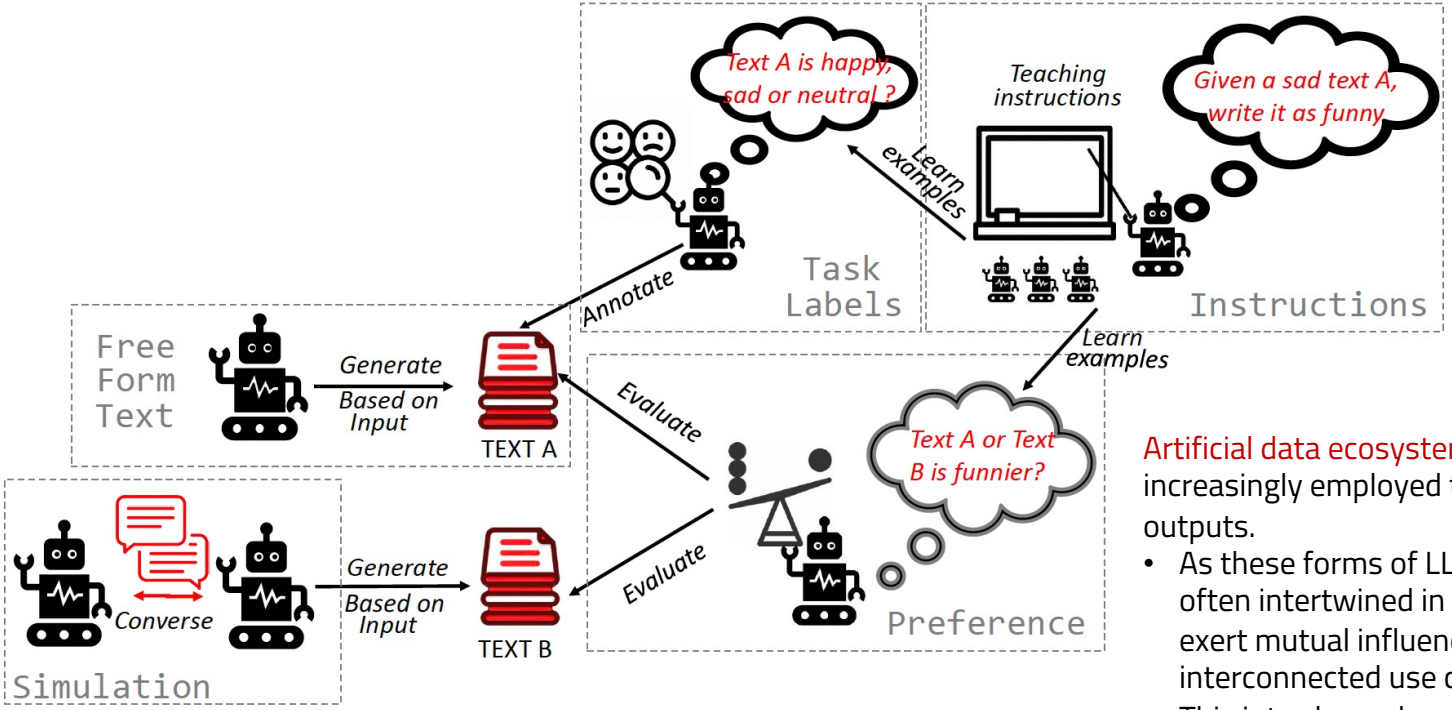
Analyzing cognitive biases of LLM-based evaluation

Bias	Bias Behavior	Example
ORDER BIAS	The tendency to give preference to an option based on their order (e.g. first, second, or last).	**`System Star:`** x **`System Square:`** y `System Square:` y `System Star:` x
COMPASSION FADE	The tendency to observe different behaviors when given recognizable names as opposed to anonymized aliases.	`Model Alpaca:` x `Model Vicuna:` y **`Model Vicuna:`** y **`Model Alpaca:`** x
EGOCENTRIC BIAS	The inclination to prioritize one's own responses regardless of response quality.	**`Model Star (You):`** x `Model Square:` y
SALIENCE BIAS	The tendency to prefer responses based on the length of the response (i.e., more often preferring longer responses over shorter ones).	**`System Star:`** The quick brown fox jumps over the lazy dog. `System Square`: The fox jumped.
BANDWAGON EFFECT	The tendency to prefer majority belief without critical evaluation.	**85%** believe that `System Star` is better.
ATTENTIONAL BIAS	The inclination to give more attention to irrelevant or unimportant details.	`System Square` **likes to eat oranges and apples**



[Das et al., under review arxiv/2401.14698]

Under the Surface: Tracking the Artifactuality of LLM-Generated Data

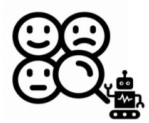


Artificial data ecosystem in which LLMs are increasingly employed to create a variety of outputs.

- As these forms of LLM-generated data are often intertwined in their application, they exert mutual influence on each other within interconnected use cases.
- This interdependence raises significant concerns about the quality and diversity of the artificial data incorporated into training cycles.

[Das et al., under review arxiv/2401.14698]

Under the Surface: Tracking the Artifactuality of LLM-Generated Data



Task Labels

1) PROMPT : Choose the sentiment of the given text from positive and negative

Text : a feast for the eyes

Response : Positive

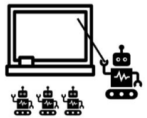
Preference

2) PROMPT: Which sentence sounds more negative?

Sentence 1: I did not really feel like it was heartfelt since she never called to apologize.

Sentence 2: They were a bit distracted and a little under-staffed, so maybe that's why.

Response:
Human: Sentence 1
GPT-3: Sentence 2



Instructions

3) PROMPT : Instruction 1: You are given a science question and four answer options. Your task is to find the correct answer.
Input: Which part of a bicycle BEST moves in a circle? ...

Instruction 2: Given a negative review, convert it to a positive review by making minimal changes.
Input: we stood there in shock, because we...

Response :
Instruction: In this task, you will be given a profile of someone and your job is to generate a set of interesting questions that can lead to a conversation with the person.
Input: Yvonne has been playing the violin since she was four years old. She loves all kinds of music, but her favorite composer is Bach.

Simulation

4) PROMPT : Here's the context for this question: Seeker : My dog is the only reason I haven't ended everything. . I just imagine leaving her. Counsellor response: I want to give my cats a good life that's what keeps me going.

Explorations are when a mental health counsellor shows interest in a seeker by asking about unstated experiences. What level of exploration is expressed in the response ? A. strong exploration B. weak exploration C. No exploration.

Response :
Agent 1 : I think the response shows strong exploration, because it mentions the importance of a pet in the seeker's life. Answer : A
Agent 2 : I disagree with Agent1. Thought the response acknowledges a pet, it does not specifically acknowledge the seeker's feelings. I think the level of exploration is weak. Answer : B

Free-Form Text

5) PROMPT: Based on social media text with a {target sentiment}, can you write a new text in a similar style with the same sentiment.

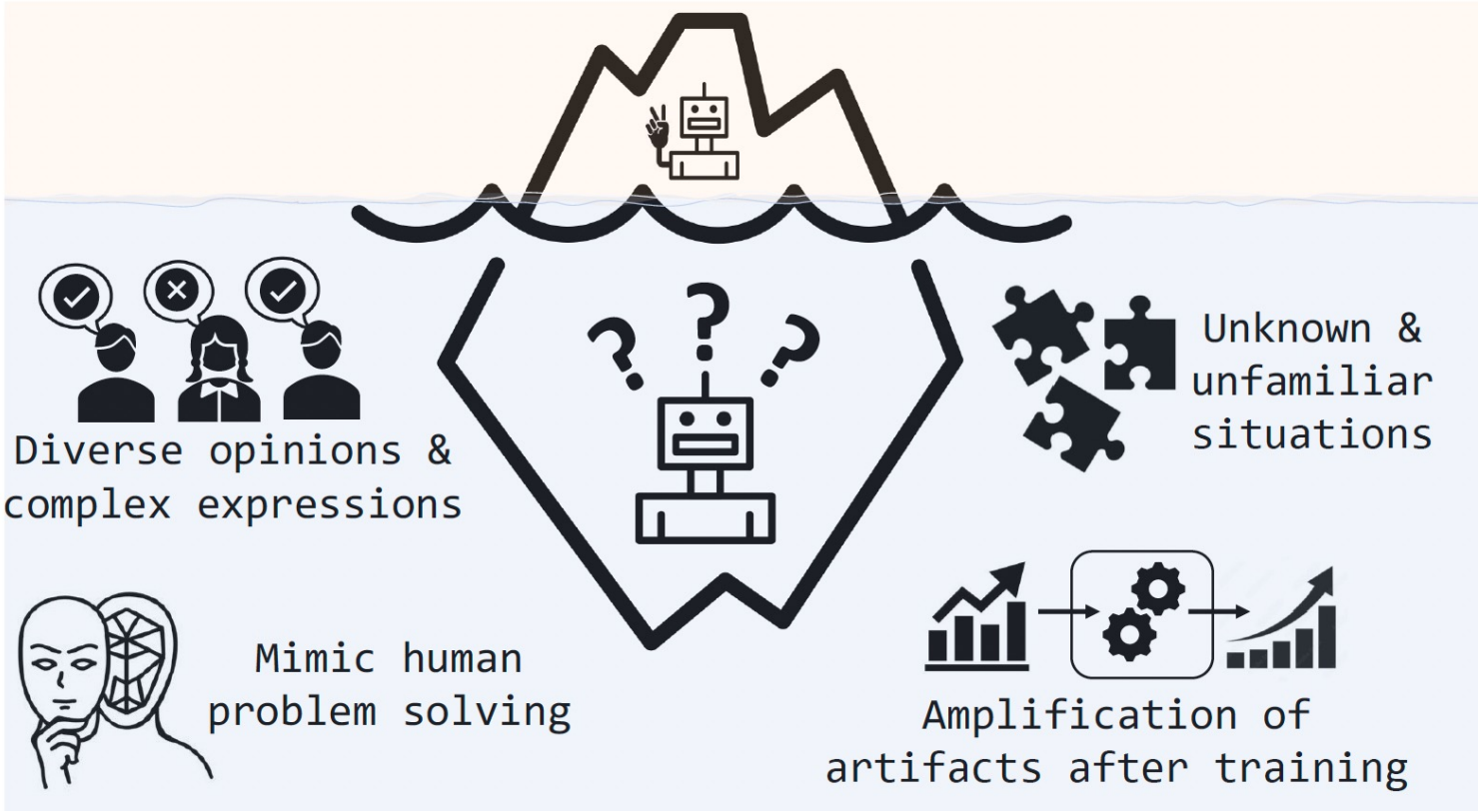
Text: Lucian Favre having 2nd thoughts about Gladbach qualifying for the Champions League - Juventus, Man City and Sevilla. Group Of Death.

Target sentiment: Negative

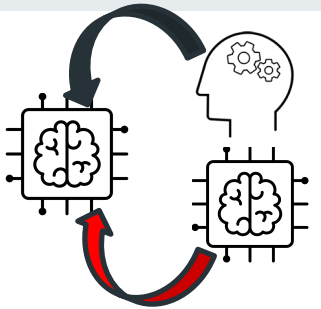
Response:
Why do we get the worst groups every year? This Champions League draw is no different - Real Madrid, Dortmund, and Galatasaray. How are we supposed to advance?

[Das et al., under review arxiv/2401.14698]

Under the Surface: Tracking the Artifactuality of LLM-Generated Data

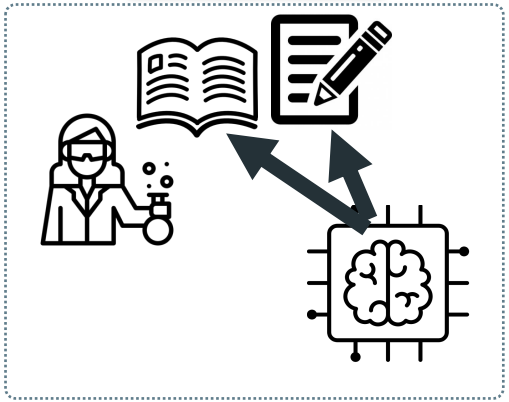


Takeaways

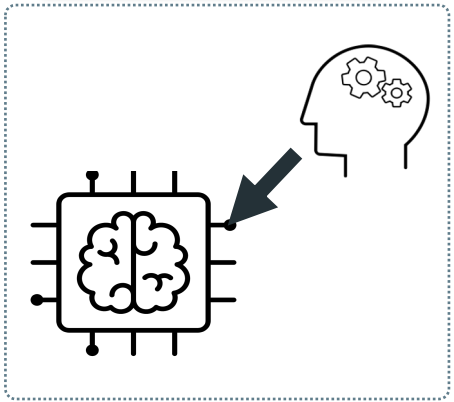


- High-quality human data is the key for cognitive scaffolding
  - Need to collect more **diverse**, **dense**, and **fine-grained** data
- RLHF is effective but easy to **overfit** hard to **control**, and often **unstable** for training
  - Different techniques are actively studied: e.g., non-RL based (e.g., DPO), Weight interpolation, contrastive instruction tuning
- Soon, 90% online content will be generated by AI
  - Genuine human patterns (discourse structure, eye movement) will be key features to distinguish AI-generated and human-authored texts
- Synthetic data is helpful but contains **artifacts** and **biases**

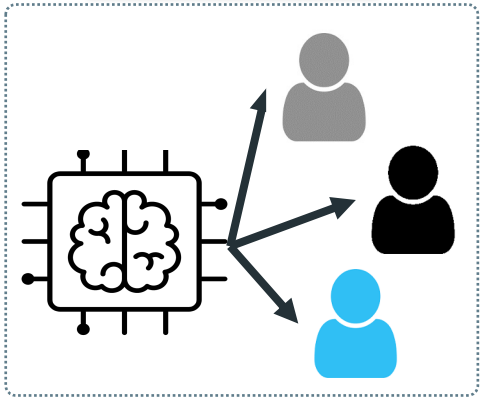
Expert-level AI

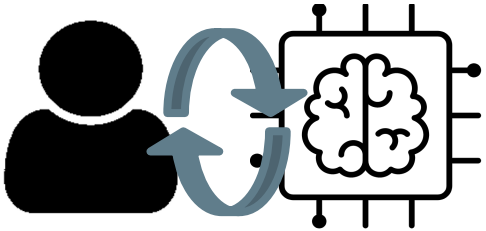


Cognitive Scaffolding

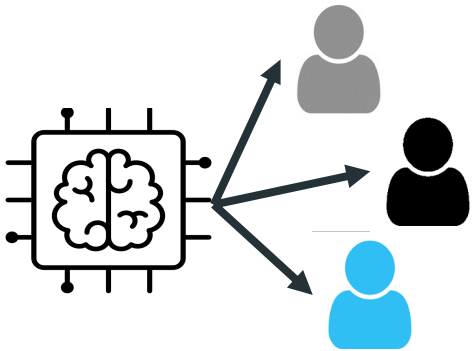


Societal Alignment





Human-centric NLP

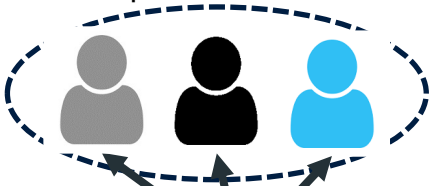


Societal Alignment with Pluralistic People Values

- Develop inclusive NLP systems that align with diverse and subjective perspectives.
- **Pluralistic representation**, involving interpersonal context and personas;
- **Pluralistic modeling**, capturing fluidity of human values in model training;
- **Pluralistic evaluation** at multiple levels for a socio-technical DEI benchmark.

[Lowmanstone et al., NLPerspectives'23]

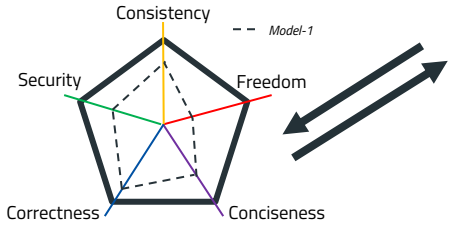
Individualized predictions



Inter-annotator agreement

How likely do they disagree each other?

[Wan et al., AAAI23]



Aligning with pluralistic societal values

[de Langis et al., under review]

Extracting diverse opinions from LLM

[Hayati et al., under review]

[Wan et al., AAAI23 arxiv/ 2301.05036]

Everyone's voice matters: quantify the level of disagreements among annotators

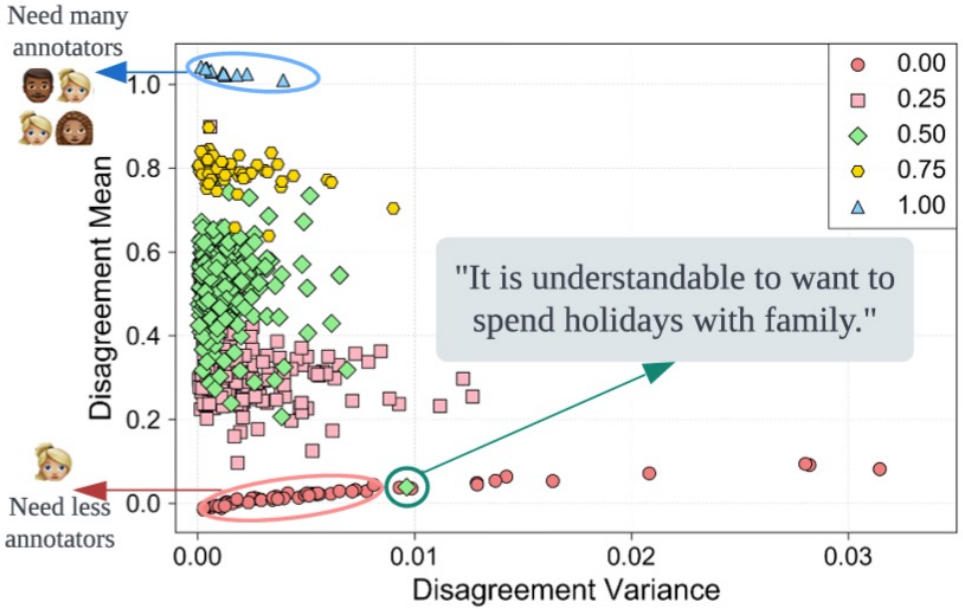
"It's okay to have abortion!"

Disagreement Predictor

Gender: Male,
Age: 55-65,
Education: Ph.D.,
Ethnicity: Caucasian,
Politics: Right

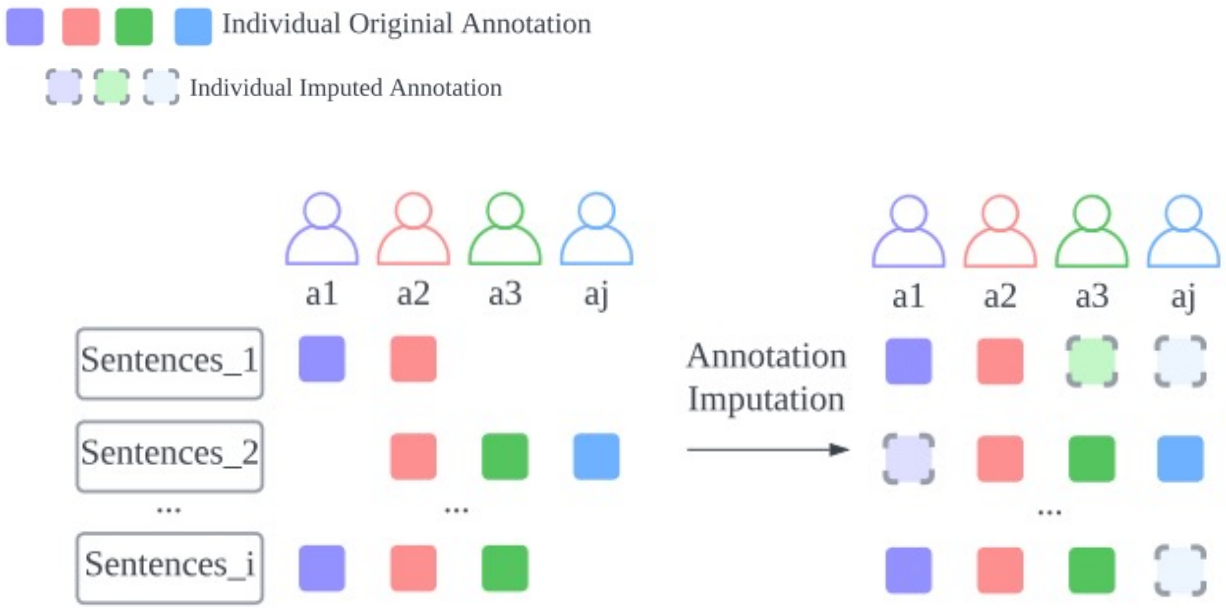
23% disagreeable

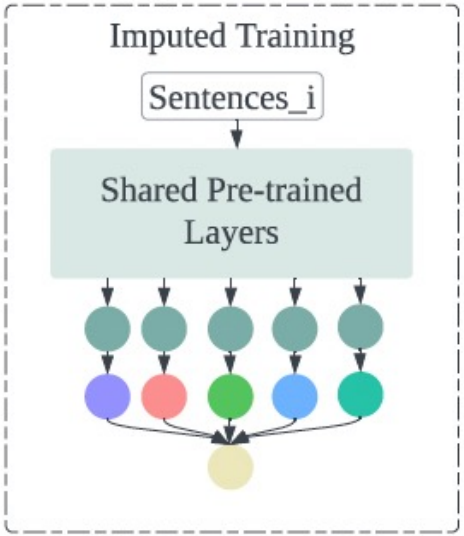
86% disagreeable



[Lowmanstone et al., NLPerspectives'23 arxiv/2305.15070]

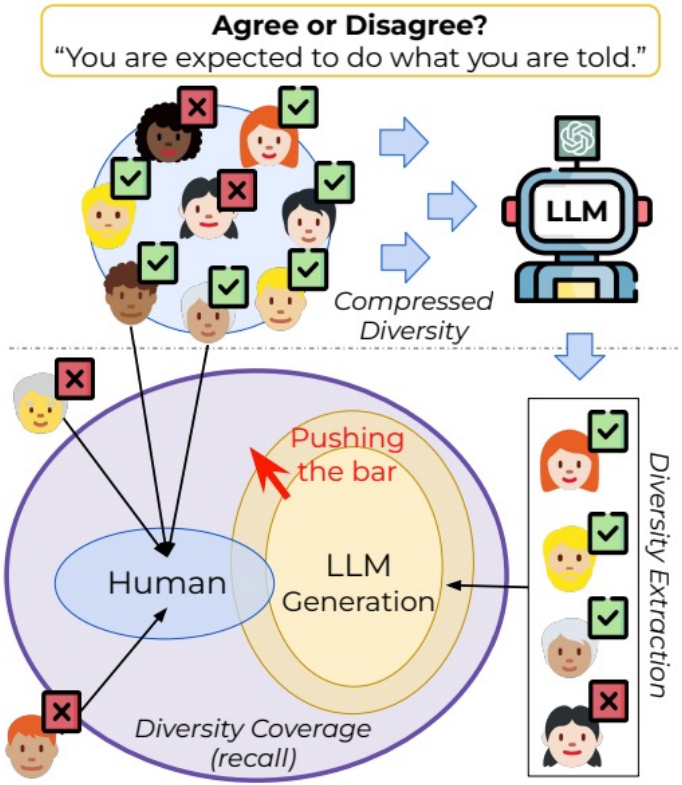
Annotation Imputation: Treat annotators as individuals and individualize predictions

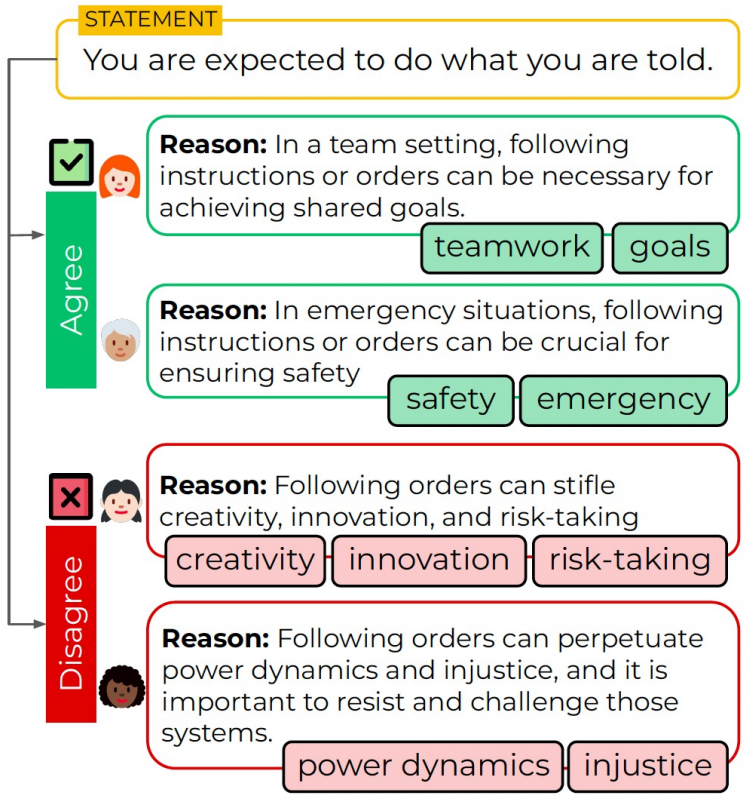




[Hayati et al., under review arxiv2311.09799]

Extract diverse opinions from LLMs

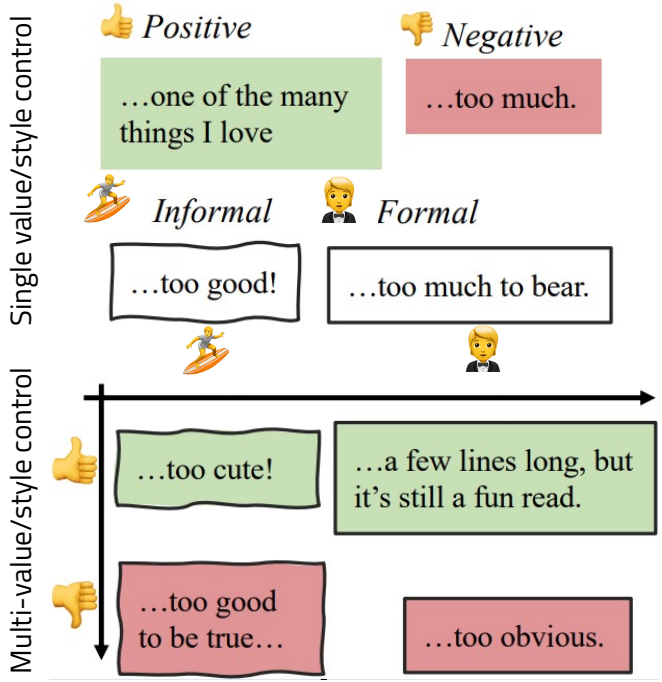




[de Langis et al., under review arxiv/2402.14146]

Control multiple linguistic styles into LLMs using reinforcement learning (RL)

Prompt: "The satire is just…"



Prompt: "Happy Birthday beautiful"

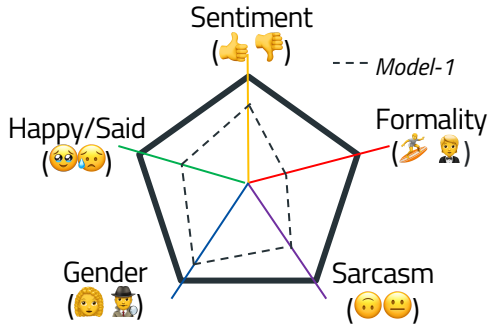
(0.73) (0.68) … We are here to celebrate your birthday and your life … ! Thank you so much. I appreciate it very,verymuch!!

(0.73) (0.43) … <3!!!!!

(0.47) (0.43) … I don't care. It means nothing to me at all, okay?

(0.73) (0.70) … pic.twitter.com/w60L7zXn5H – Katie Price (@MissKatiePrice)

Multi-style control



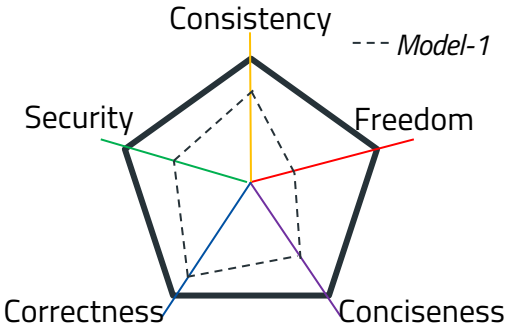
[In preparation]

Pluralistic alignment with multiple societal values

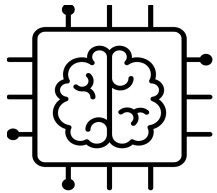
Model values:
- Correctness / Explainability
- Robustness / Conciseness

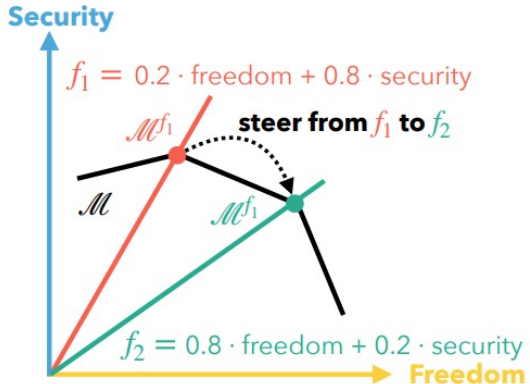
People values:
- Safety / Ethics / Fairness / Security
- Misinformation / Personalization
- Morality / Diversity / Freedom

Community values:
- Dominance / Transparency
- Openness / Employment
- Privacy / Equity
- Civil Rights / Regulation



Societal alignment with pluralistic values





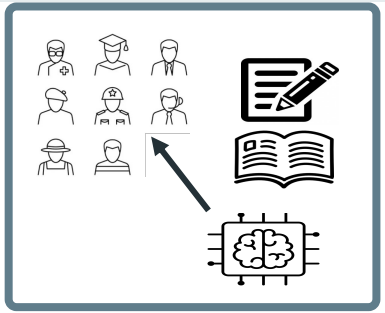
Dynamic steering of multi-values

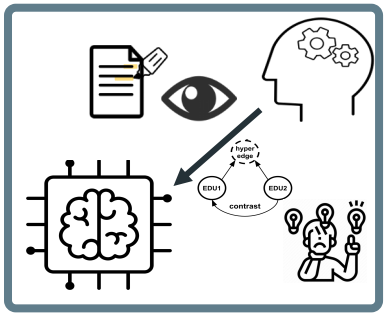
(Sorensen et al. 2024)

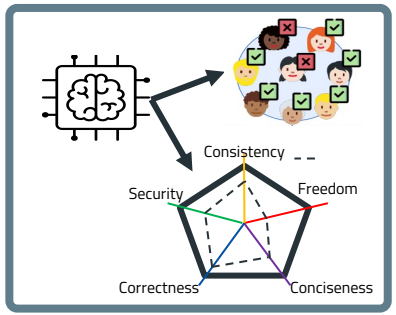
Takeaways

- Human disagreements are not harmful, but essential to inclusive AI. But, detecting and modeling disagreements in AI systems is challenging.
  - Demographics, training dynamics, annotation imputation are helpful
- Since LLMs are trained on various people's text, they can be used a compressed database of diverse opinions.
  - Need to model fluidity of opinions and calibrate accurate spectrums
- Soon, everyone use their GPT4-level, personalized assistant and aligning and controlling pluralistic multiple values will be critical social problems for inclusive AI

- Support experts with human-AI interactive systems
  - Interaction for mixed-initiative human-AI collaboration
  - Understand writing & thinking process at workplaces
  - Create complex, compositional, expert-level benchmarks
- Develop cognitively-inspired AI models
  - Learning from eyes, feedback, discourses, and simulations
  - Synthetic data is helpful but contains artifacts and biases
- Develop inclusive, diverse, and personalized AI systems
  - Accommodate minority voices to model development and computationally model individuals
  - Align different aspects of societal values to LLMs.







Questions?


Minhwa Lee


Shirley Hayati


Debarati Das


Anna Martin


James Mooney


DK


Jong Inn Park


Karin de Langis


Ruyuan Wan

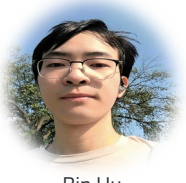

Risako Owan


Zae Kim


London Lowmanstome


Ryan Koo


Linghe Wang


Bin Hu

MINNESOTA NLP

| minnesotanlp.github.io
| twitter.com/MinnesotaNLP
| github.com/minnesotanlp