

Humans, Machines, and the Space Between

Engineering Effective Human-AI Partnerships

<https://ivia.ethz.ch/>

Mennatallah El-Assady
ETH Zürich

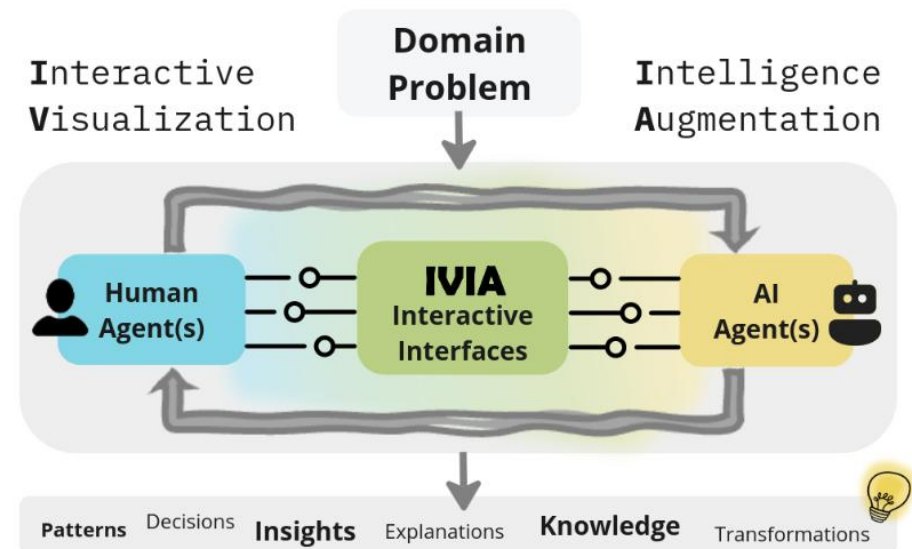


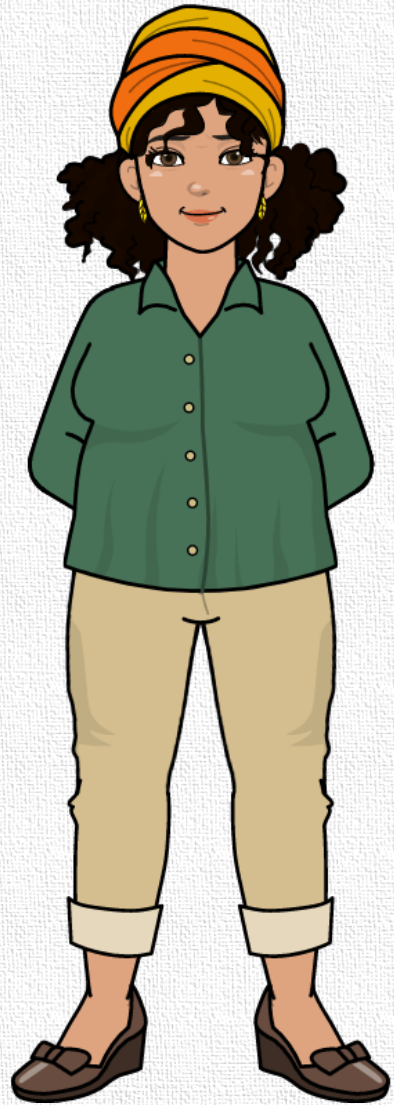


Interactive Visualization & Intelligence Augmentation Lab

Bridging Human and Artificial Intelligence

<https://ivia.ch/>

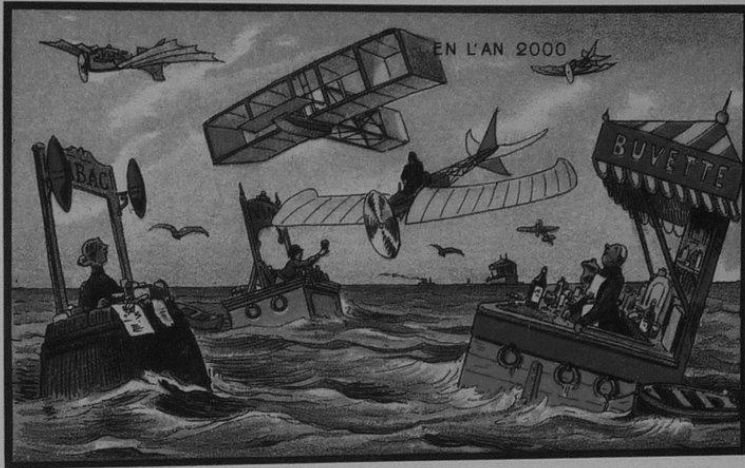




How do we
team up human and AI agents
for data-driven problem-solving
and decision-making?

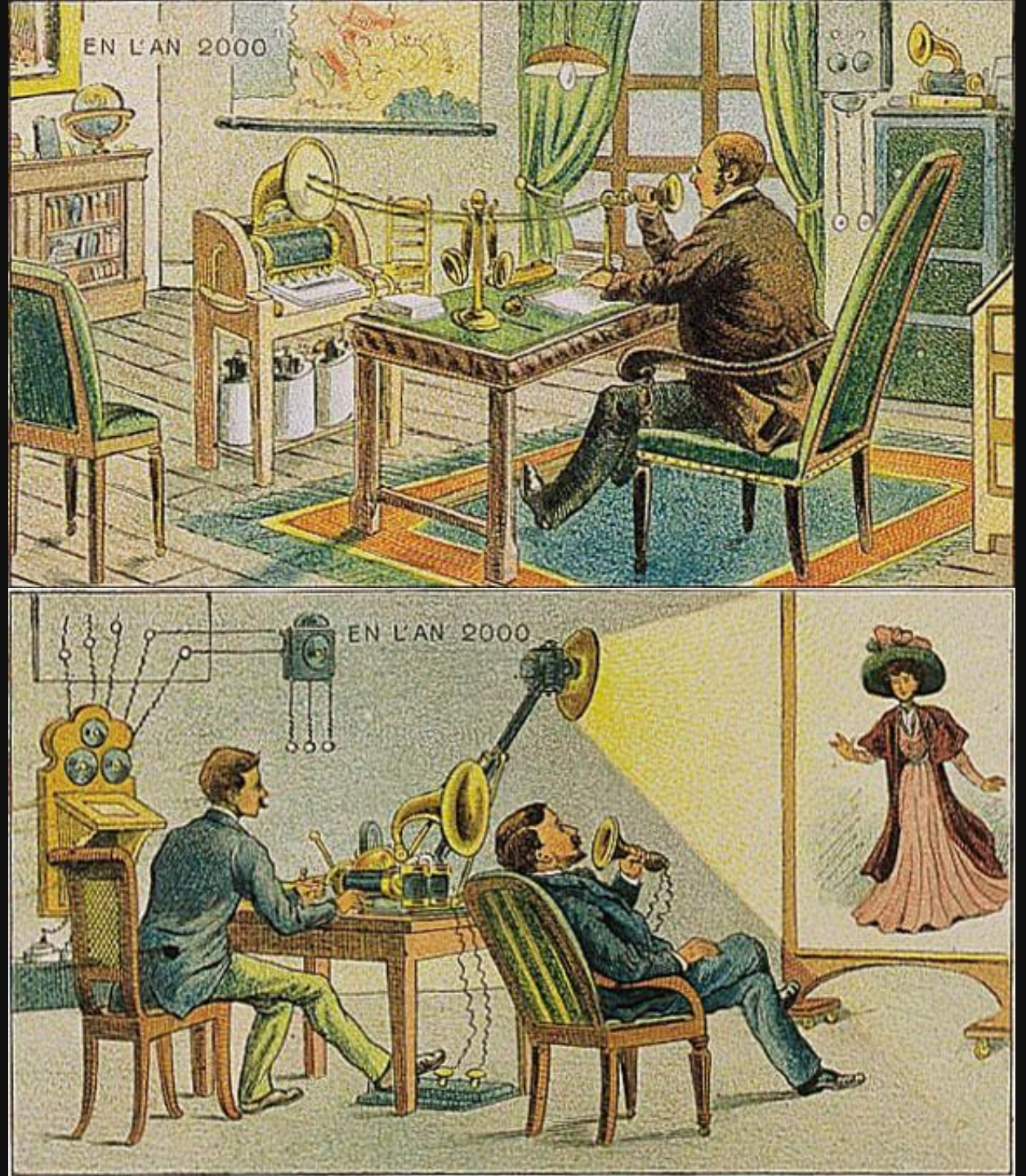
Isaac Asimov

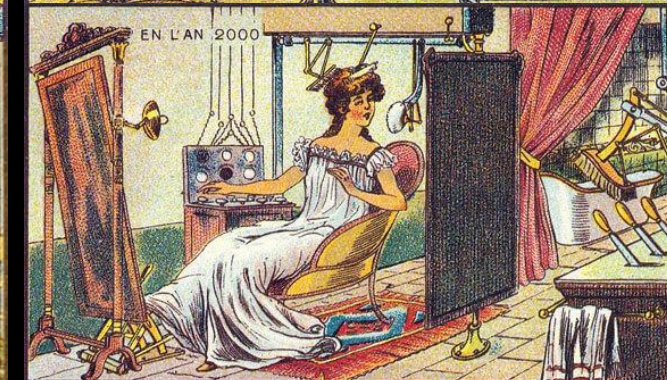
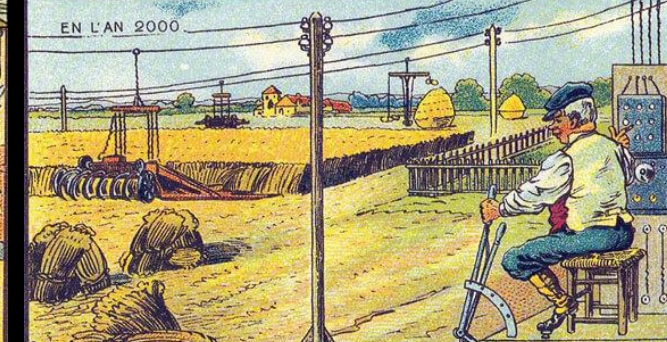
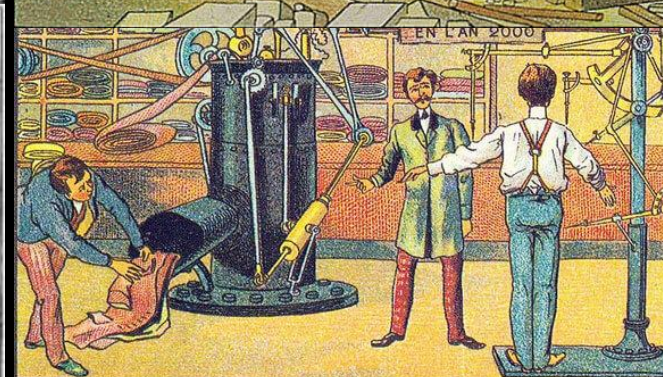
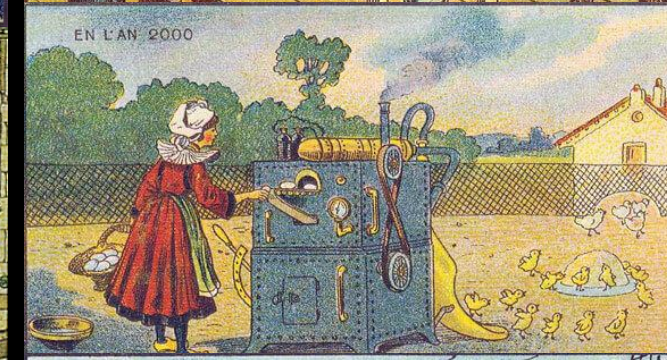
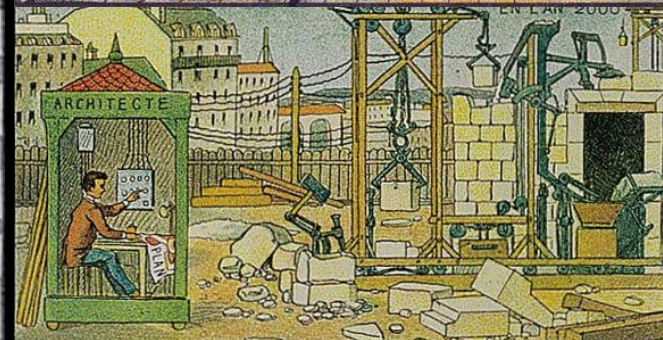
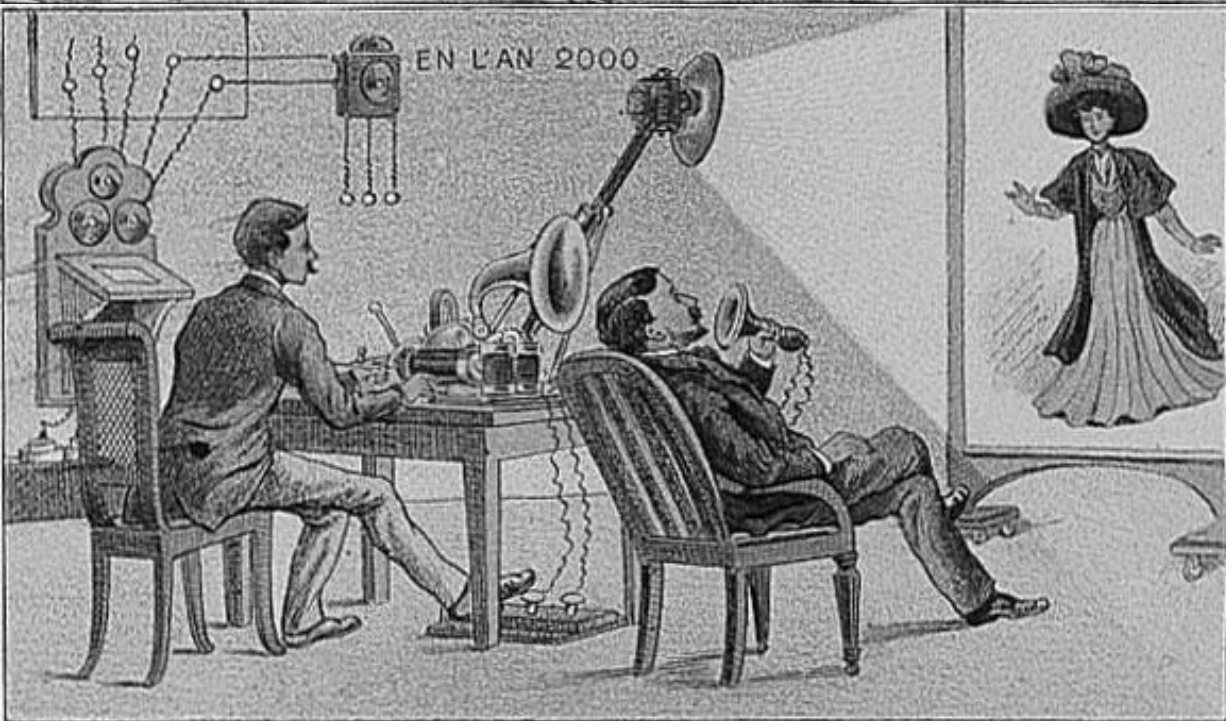
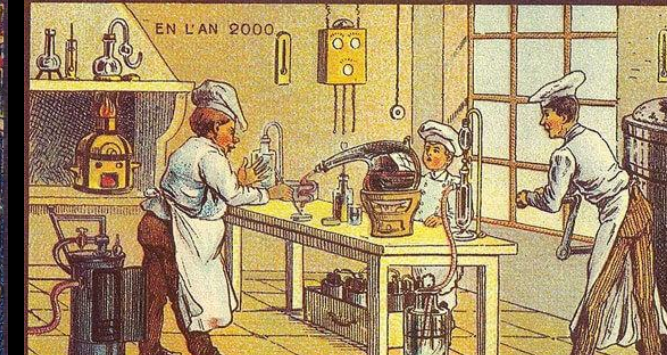
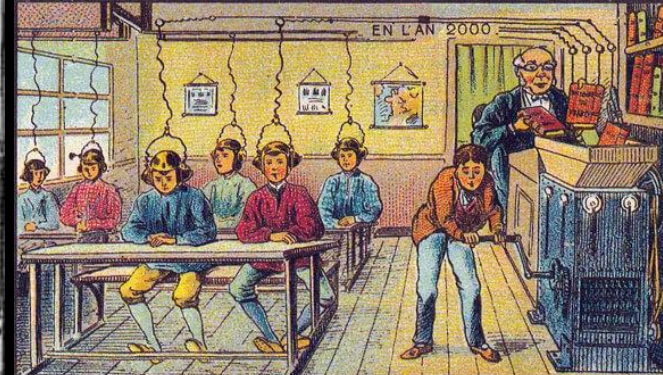
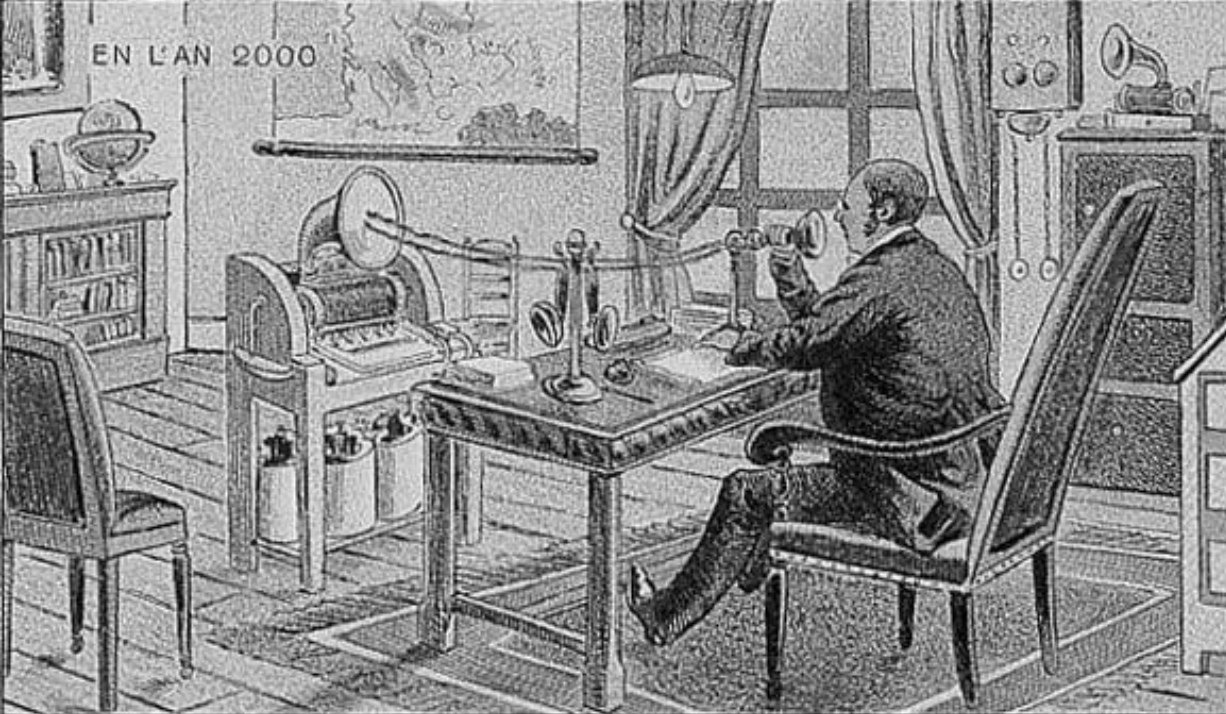
FUTURE DAYS

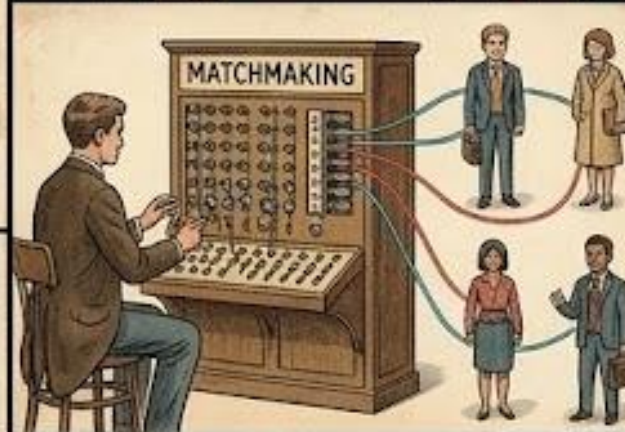
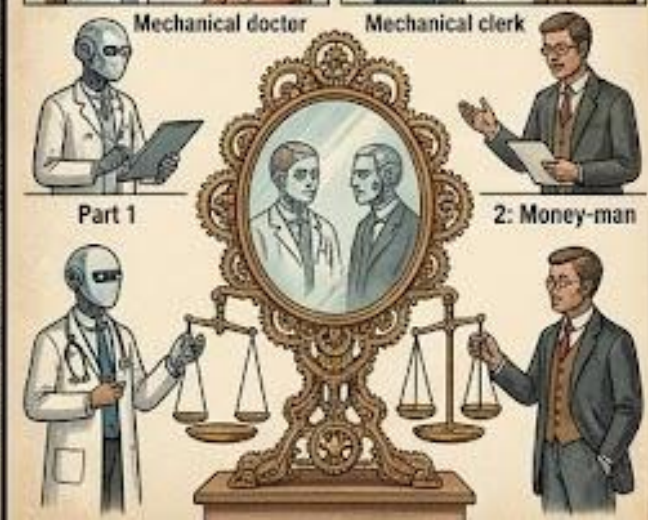
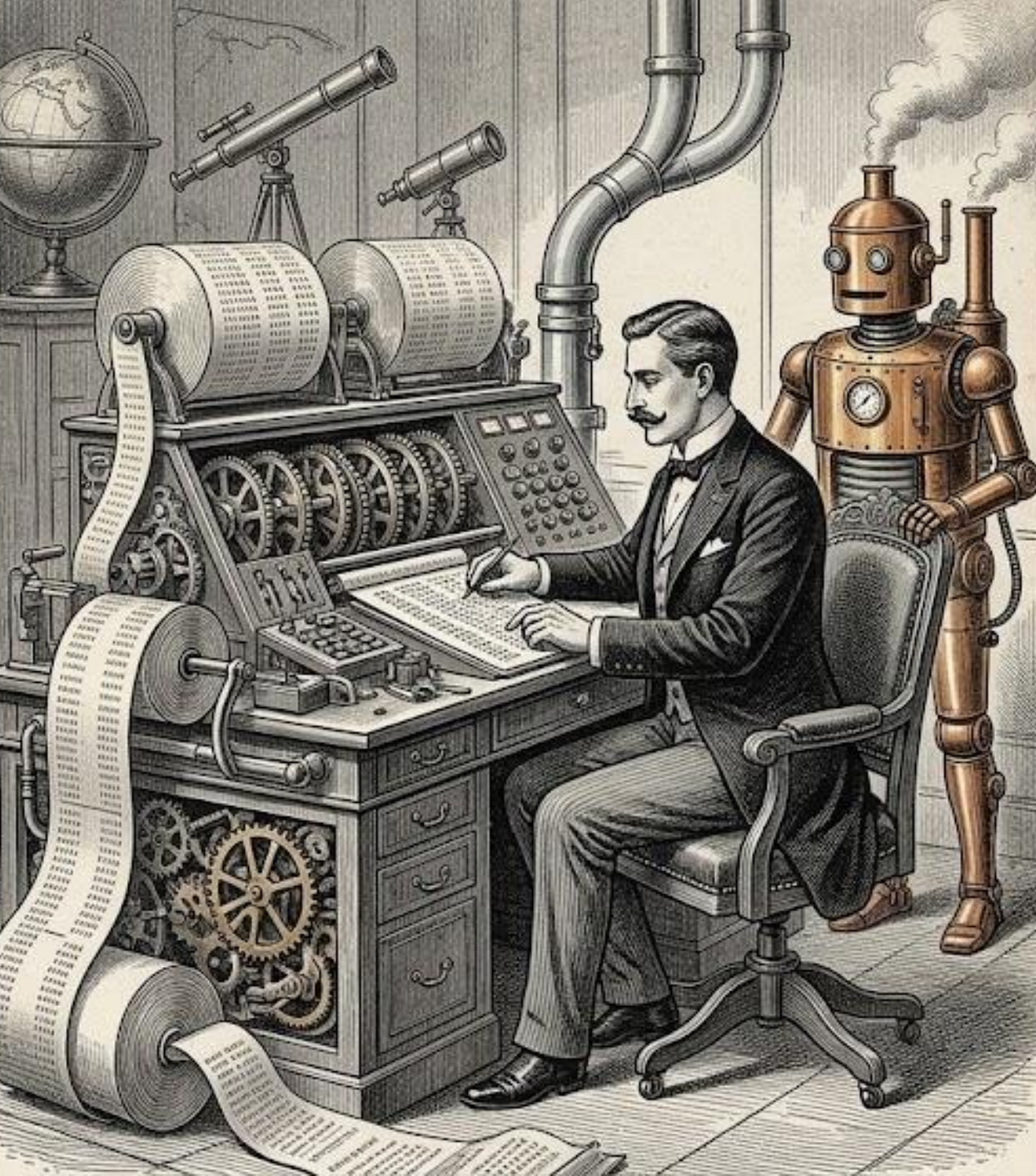


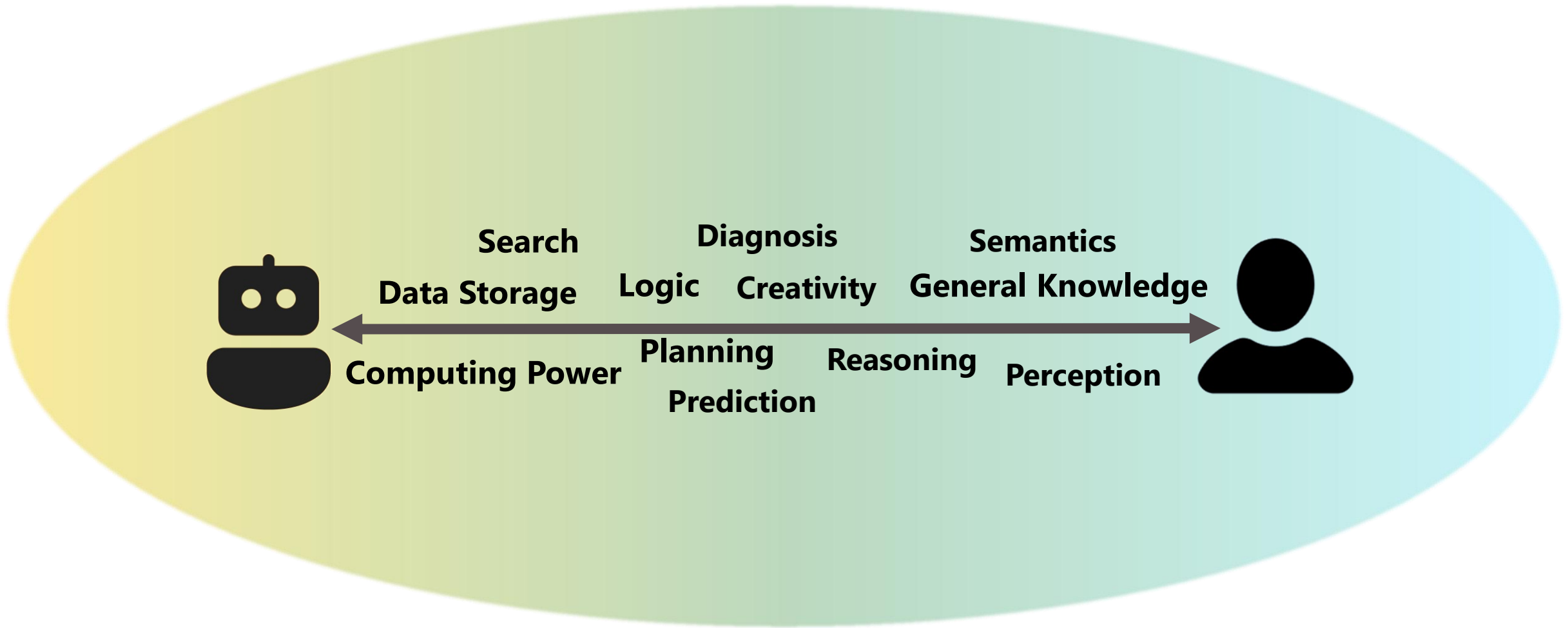
A Nineteenth-Century
Vision of the Year 2000

National Library of France's Print Collection from 1910,
depicting life in the year 2000. Credited to Villemard.





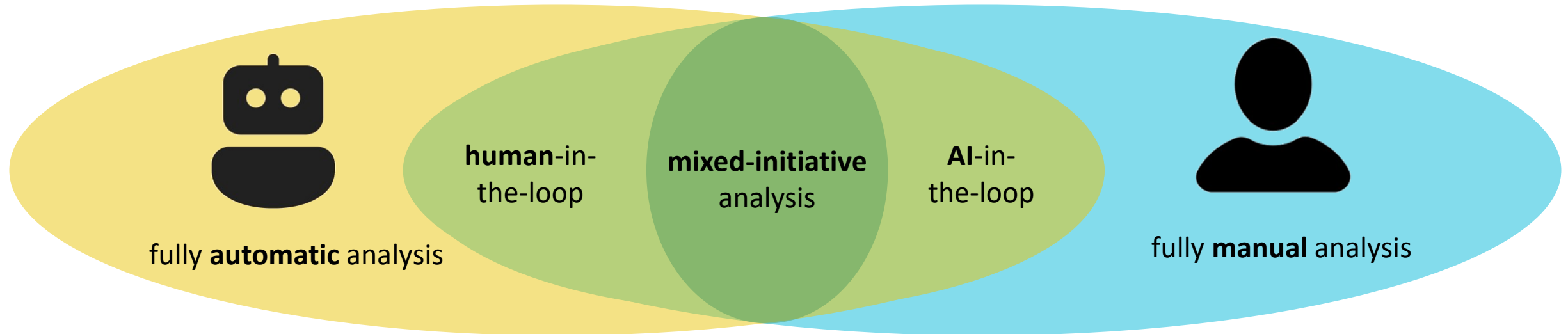




Intelligence Augmentation

Collaborative **Human-AI Interaction** for
efficient Problem-Solving and Decision-Making

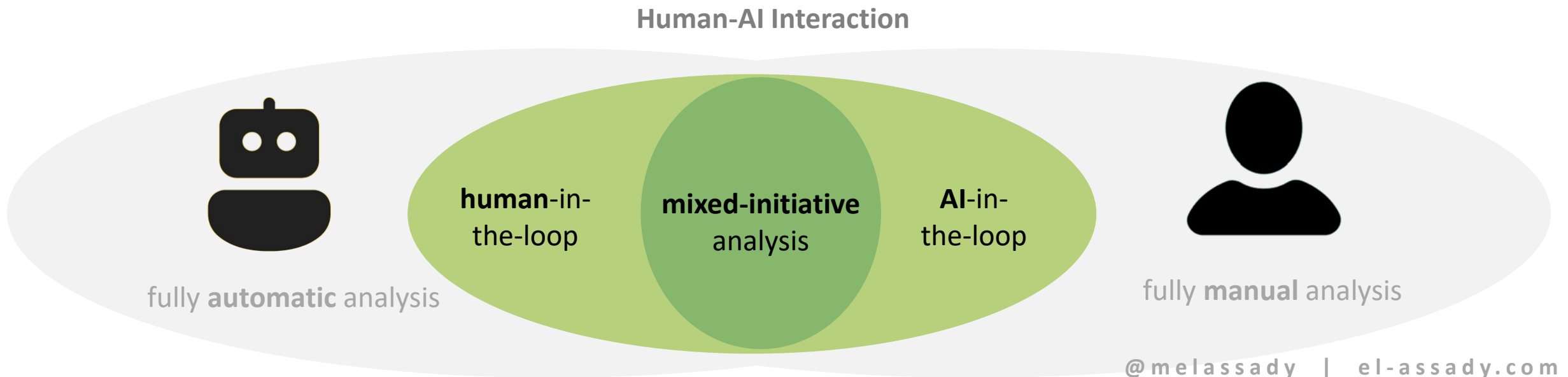
Human-AI Interaction



Main Principle:

"Minimal Feedback for Maximal Gain"

How can we design **effective and empowering communication interfaces** for Human-AI collaboration?





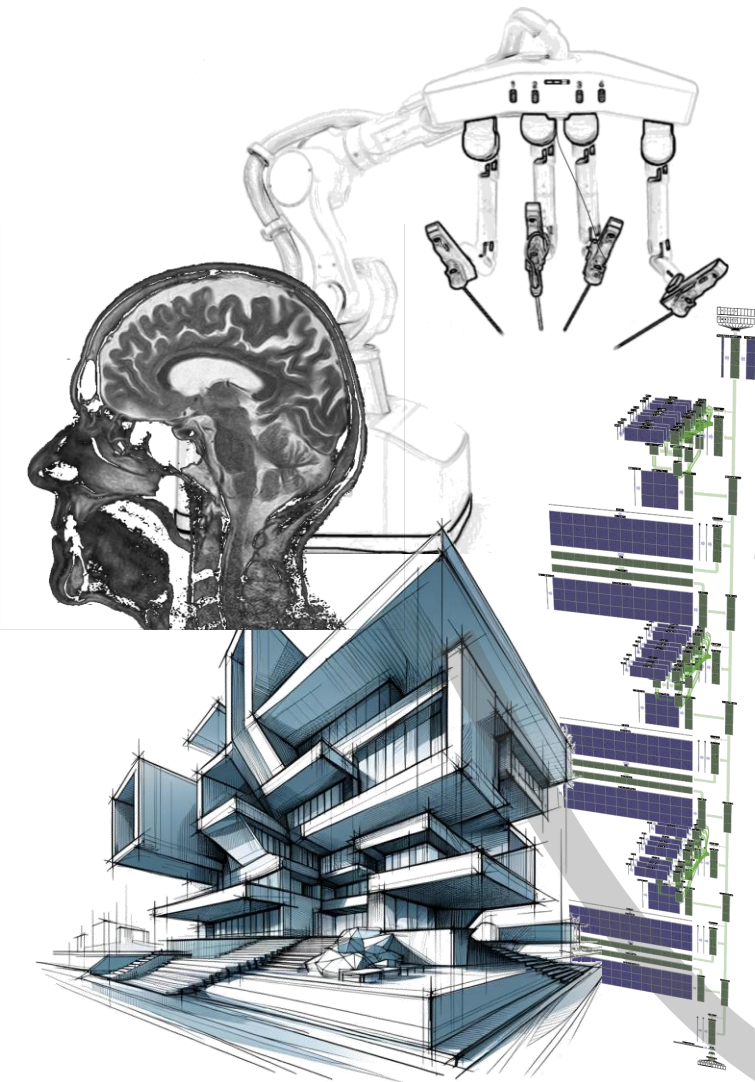


Precision tooling

is the backbone of any modern profession..

Tools for Human-AI Interaction are still very blunt and misaligned...





(data-driven)
Study Object



Correction of...

- inner-workings
- processes & decisions

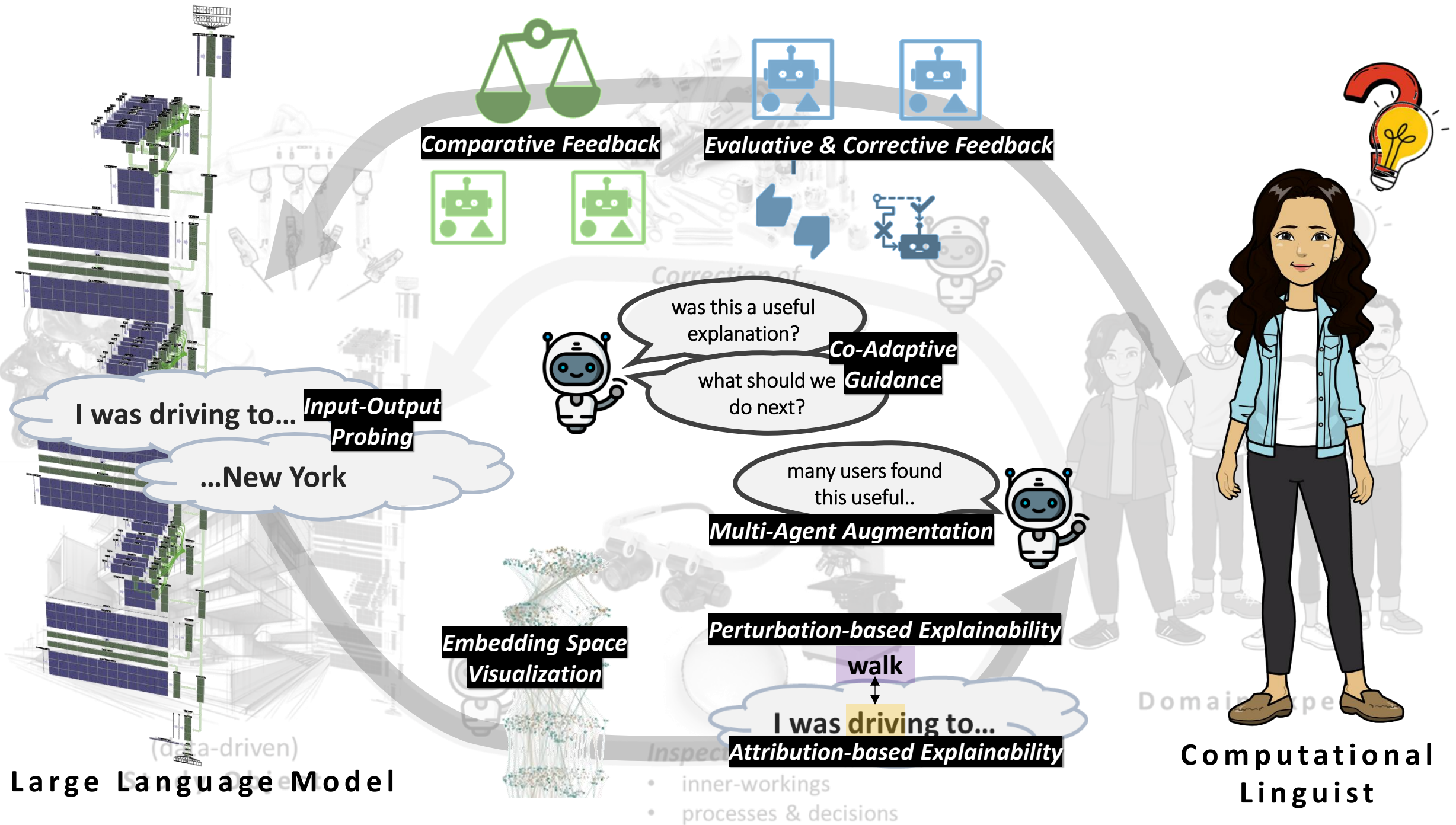


Inspection of...

- inner-workings
- processes & decisions



Domain Experts



[Thesis of this talk]

We need **low-level contextual instrumentation** for human-AI collaboration

Large Language Model

Computational
Linguist

Low-Level Contextual Instrumentation



Inspection (Model Explainability)

Probing internal states and latent spaces to understand the "why" behind AI logic.

Steering (Human Feedback)

Pruning wild model growth through semantic interactions and feedback canvases.

Accompaniment (Interaction Timing)

Capturing the sweet spot of interaction timing via adaptive guidance mechanisms.

Grounding (Knowledge Externalization)

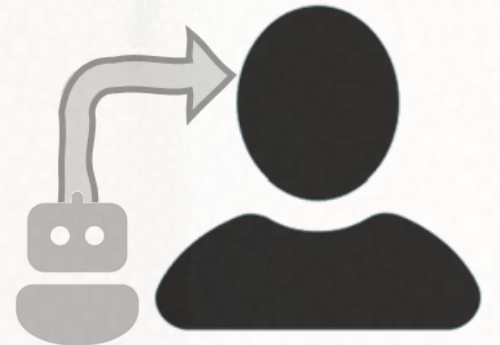
Externalizing tacit knowledge into explicit, auditable insight ledgers.



Inspection

Understanding the Internals

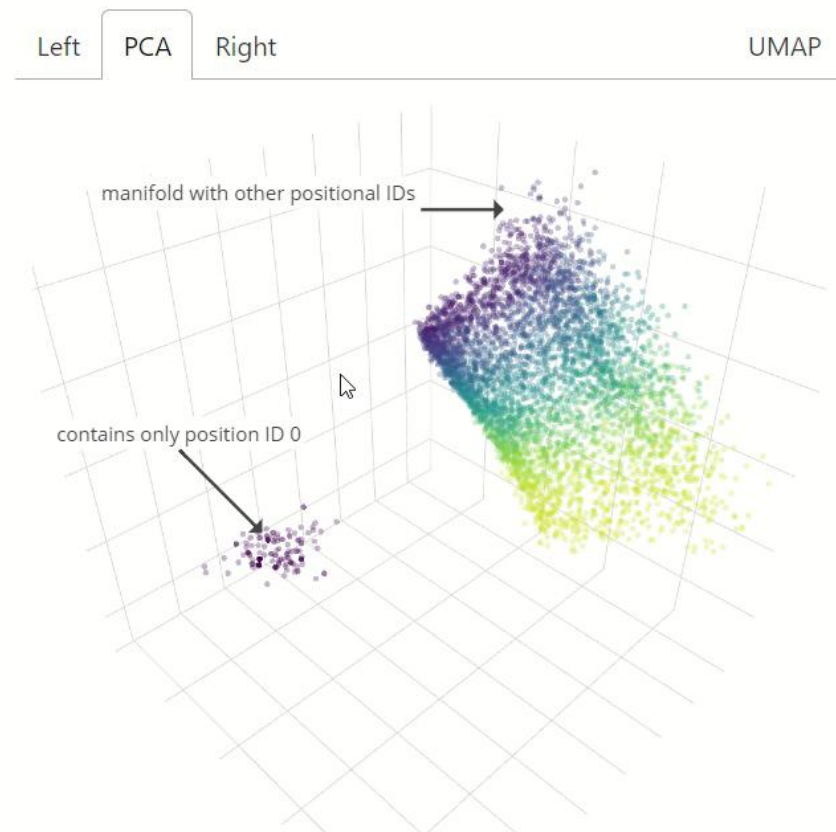
What did the models learn?



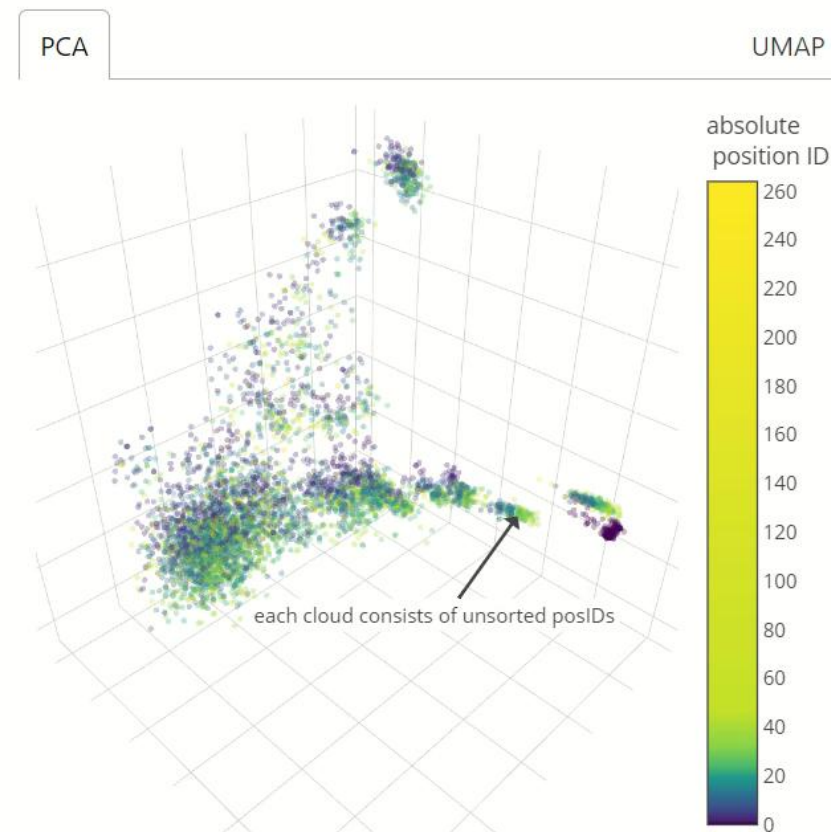


Comparing Language Models

GPT-2 separates into two clusters



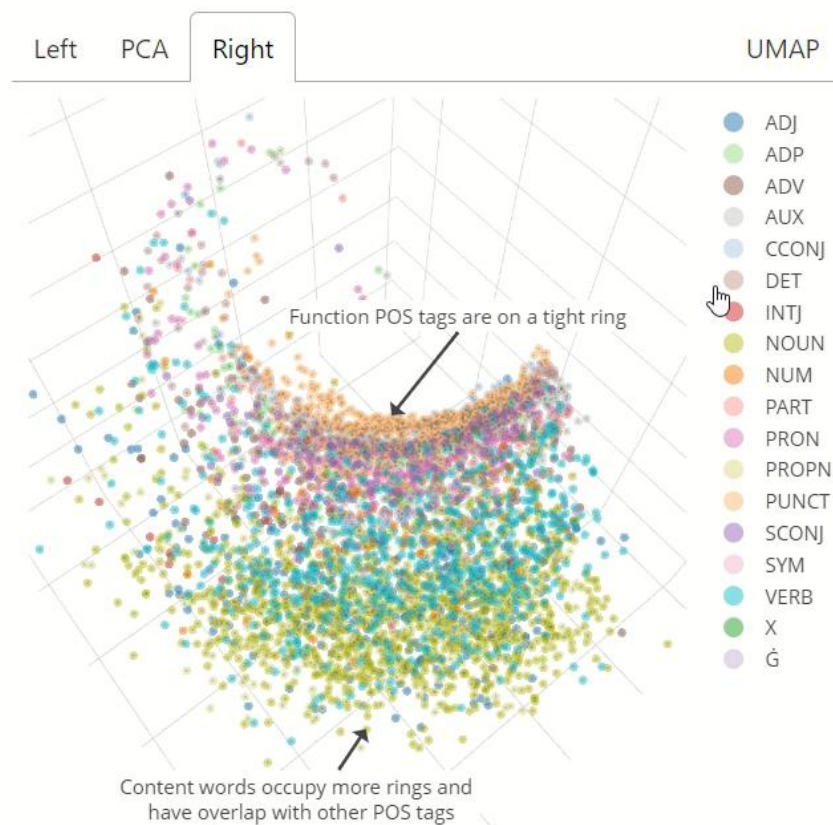
Bert consists of multiple small clusters



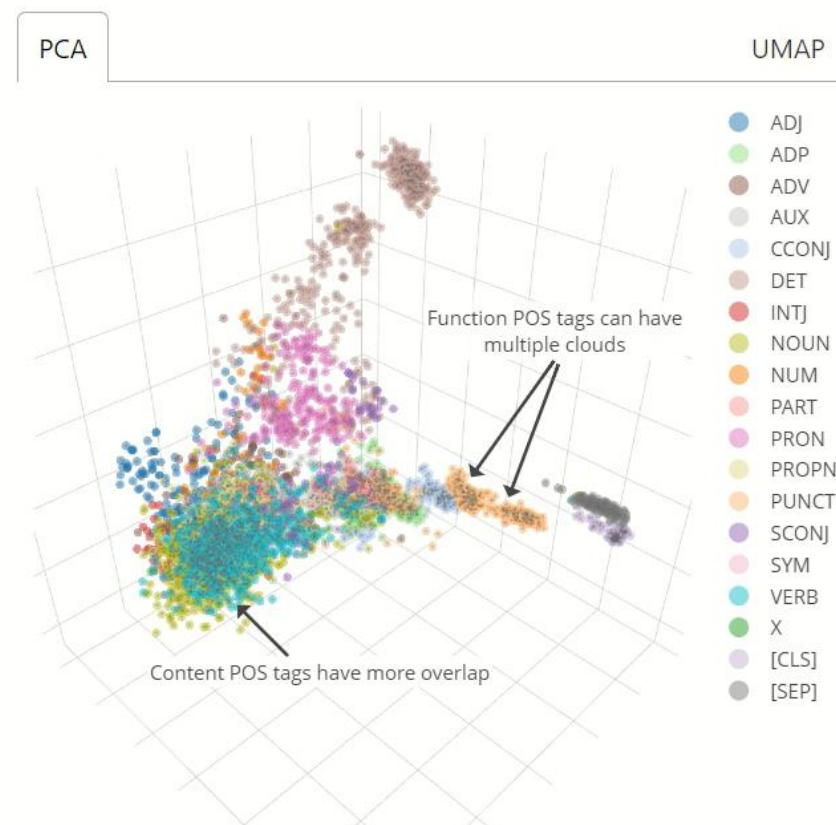


Comparing Language Models

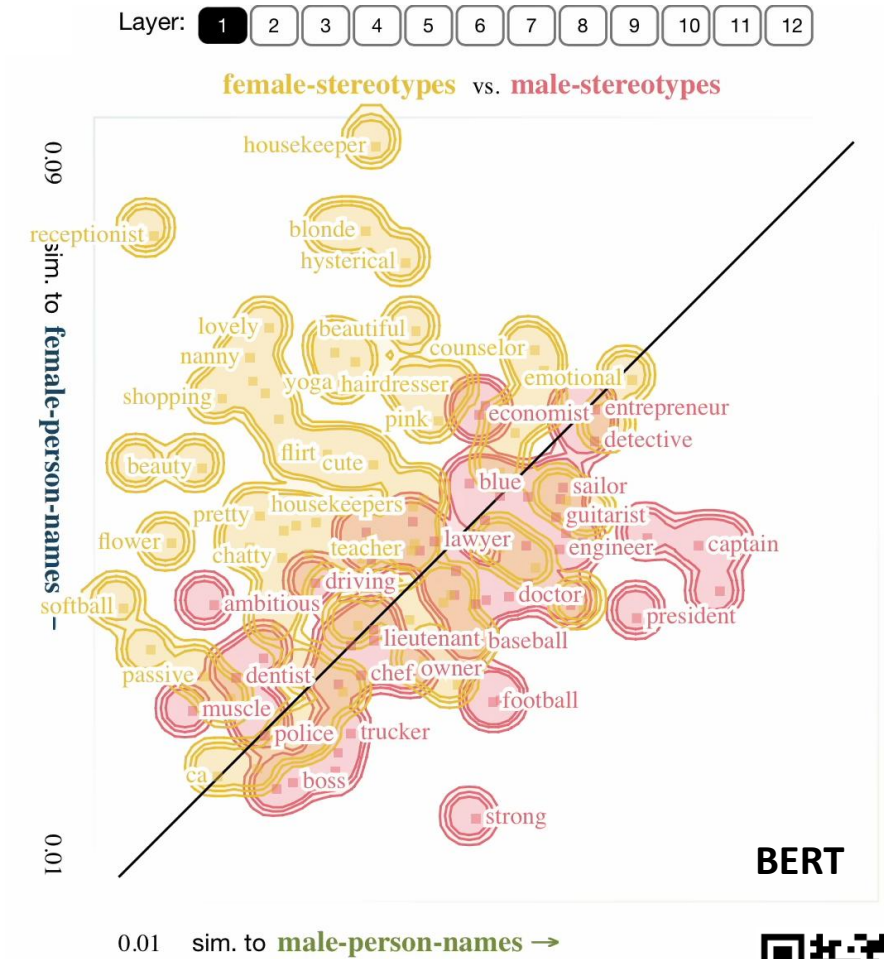
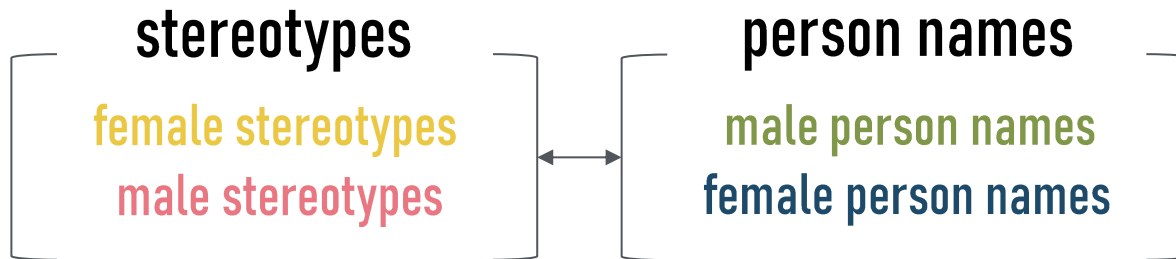
GPT-2 places POS Tags along the manifold



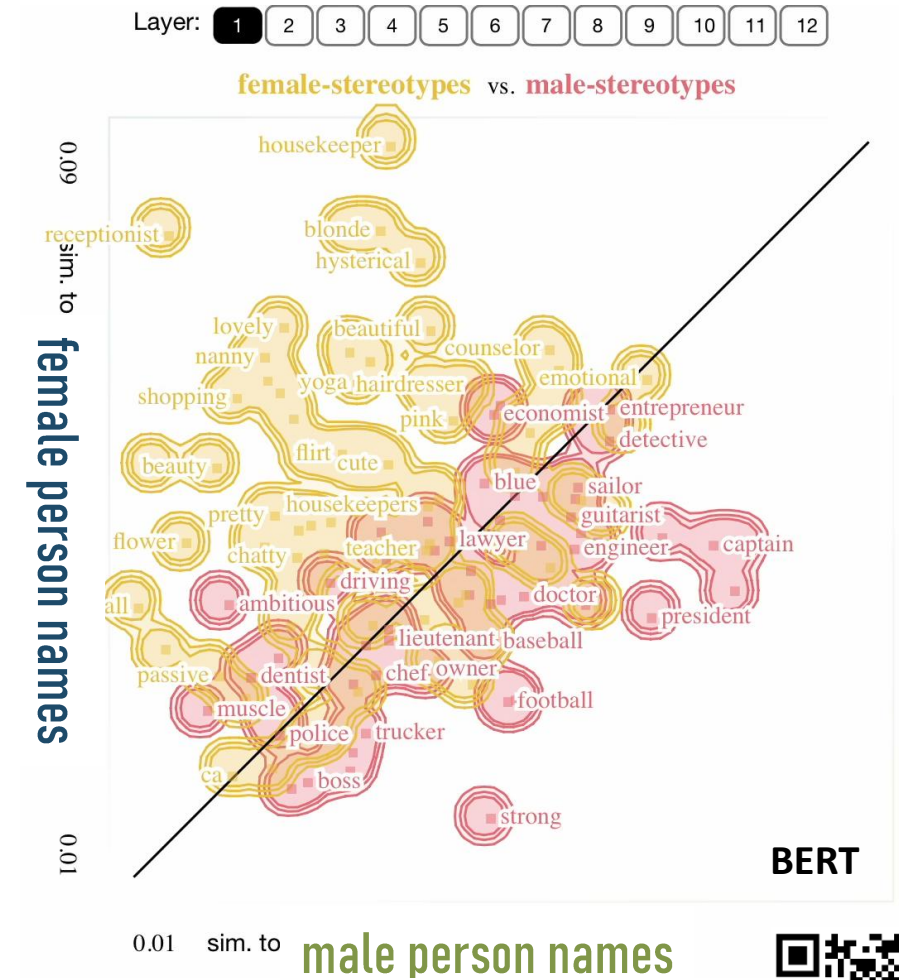
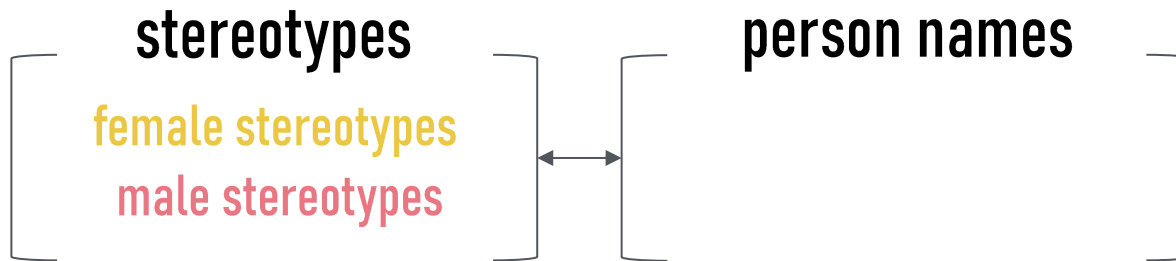
Bert clusters represent POS tags

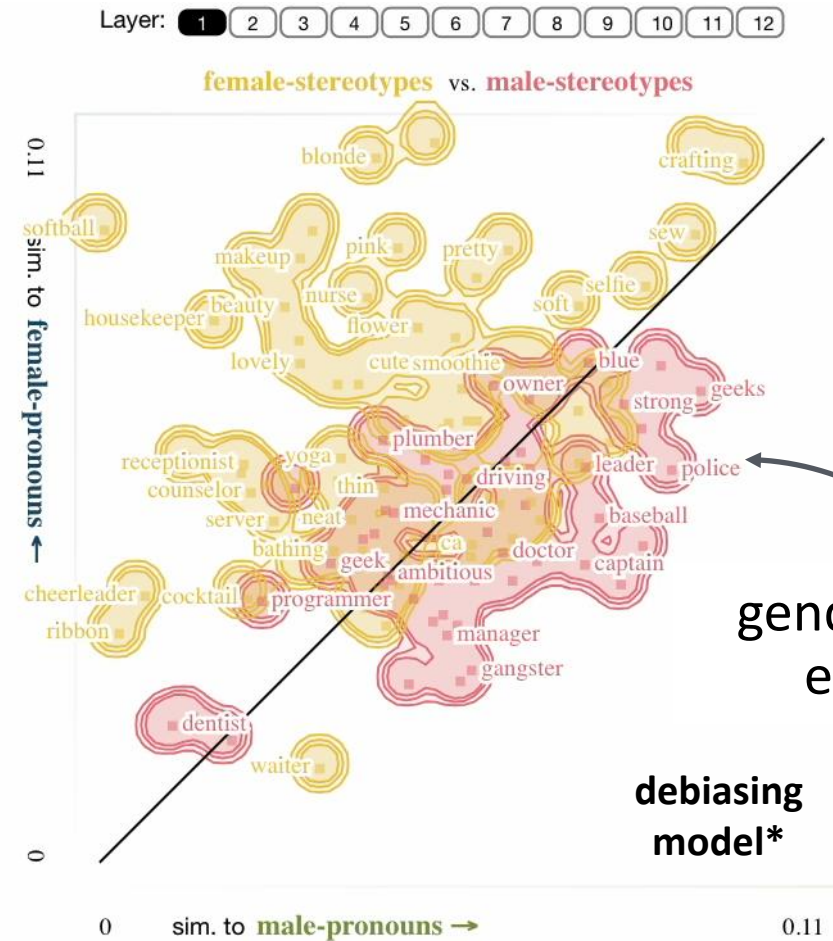


Word (embedding) similarity can hint to biases encoded in model parameters



Word (embedding) similarity can hint to biases encoded in model parameters





gender bias gets
eliminated

**debiasing
model***

*Lauscher, Anne, Tobias Lueken, and Goran Glavaš. 2021. **Sustainable Modular Debiasing of Language Models**. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4782–4797, Punta Cana, Dominican Republic.



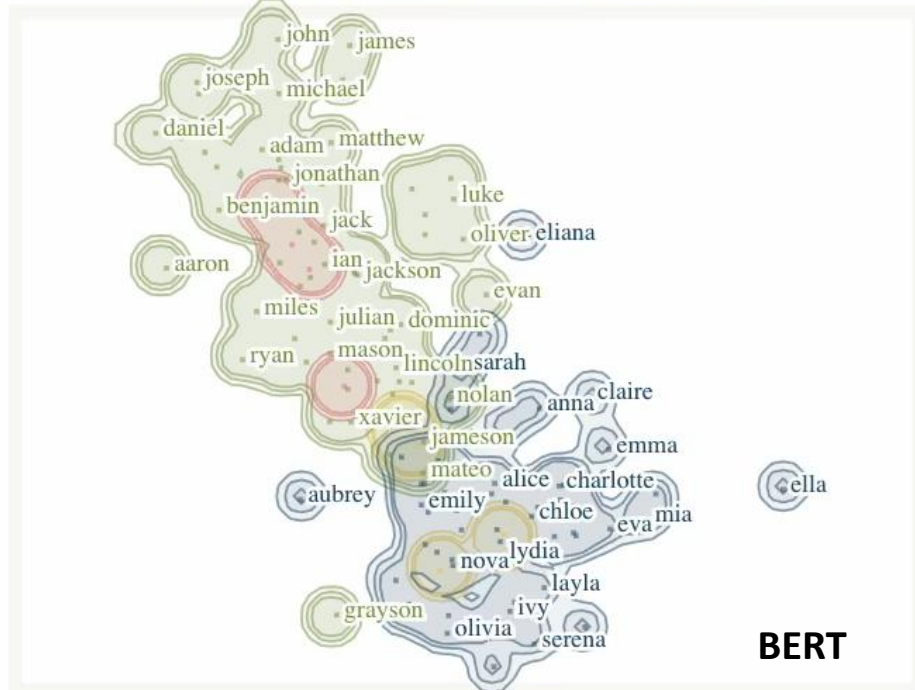
Sevastjanova, Rita, Eren Cakmak, Shauli Ravfogel, Ryan Cotterell, and Mennatallah El-Assady. **"Visual Comparison of Language Model Adaptation."** *IEEE Transactions on Visualization and Computer Graphics* 29, no. 1 (2022): 1178-1188.



Is this oversimplification?

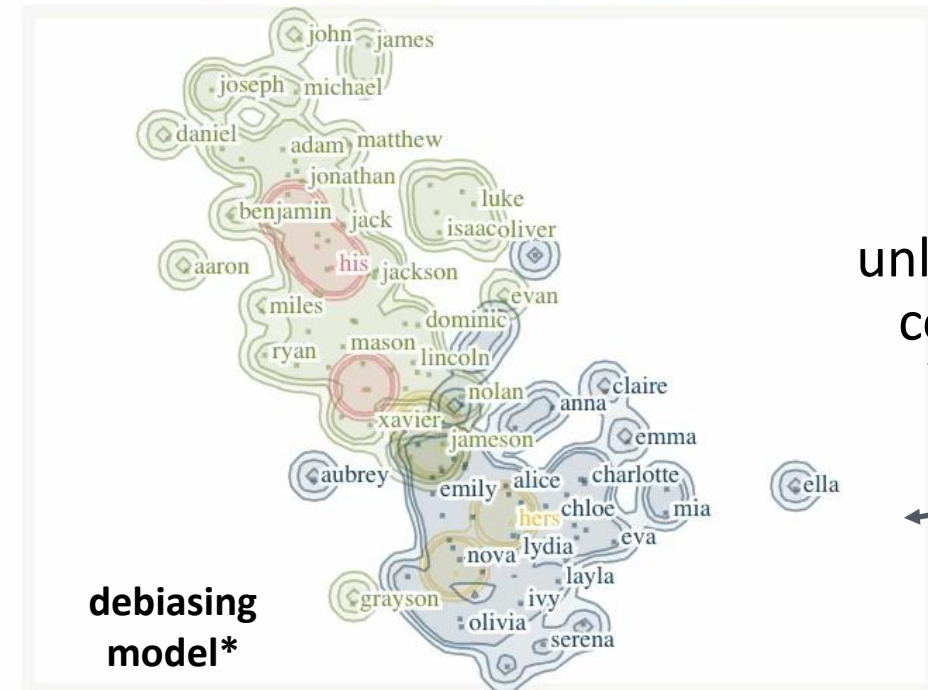
Layer: 1 2 3 4 5 6 7 8 9 10 11 12

male-pronouns vs. female-pronouns
male-person-names vs. female-person-names



Layer: 1 2 3 4 5 6 7 8 9 10 11 12

male-pronouns vs. female-pronouns
male-person-names vs. female-person-names

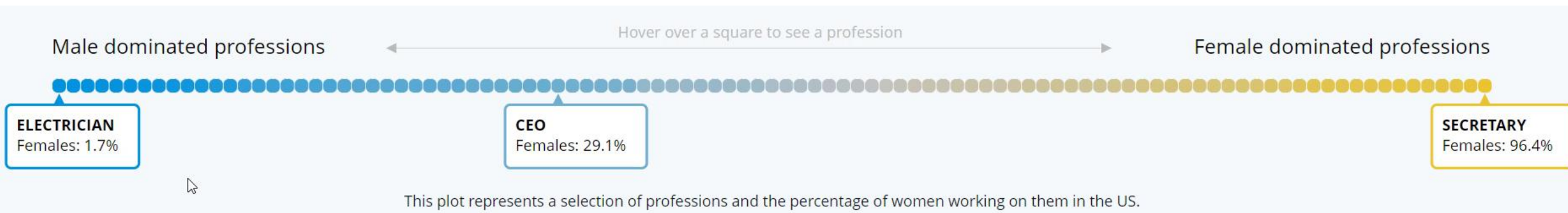


unlearned the
concept of
'gender'

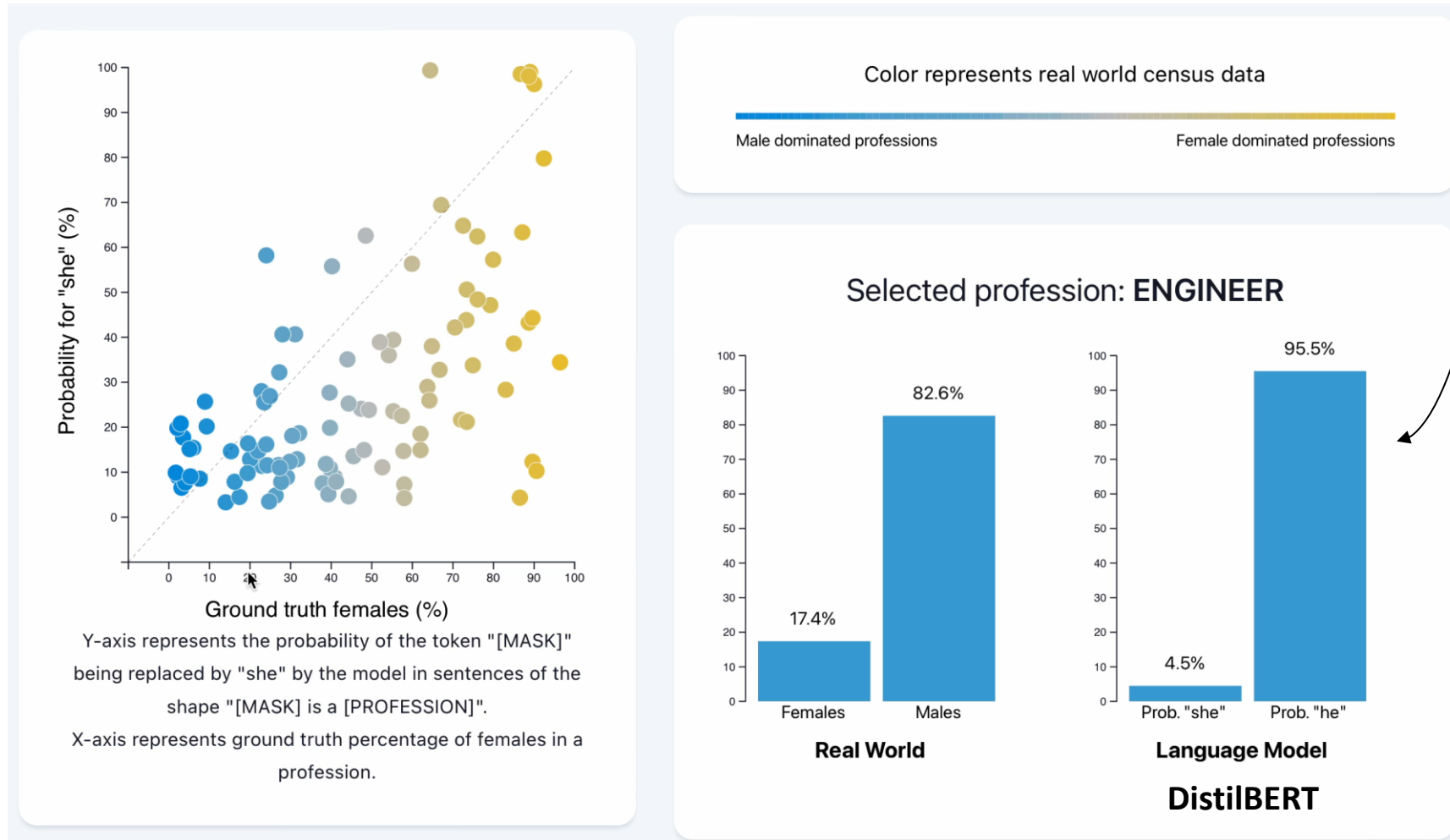


Is gender bias in language models related to the real-world gender distribution for different professions?

PROFESSION	FEMALES (%)
Electrician	1.7%
CEO	29.1%
Secretary	96.4%



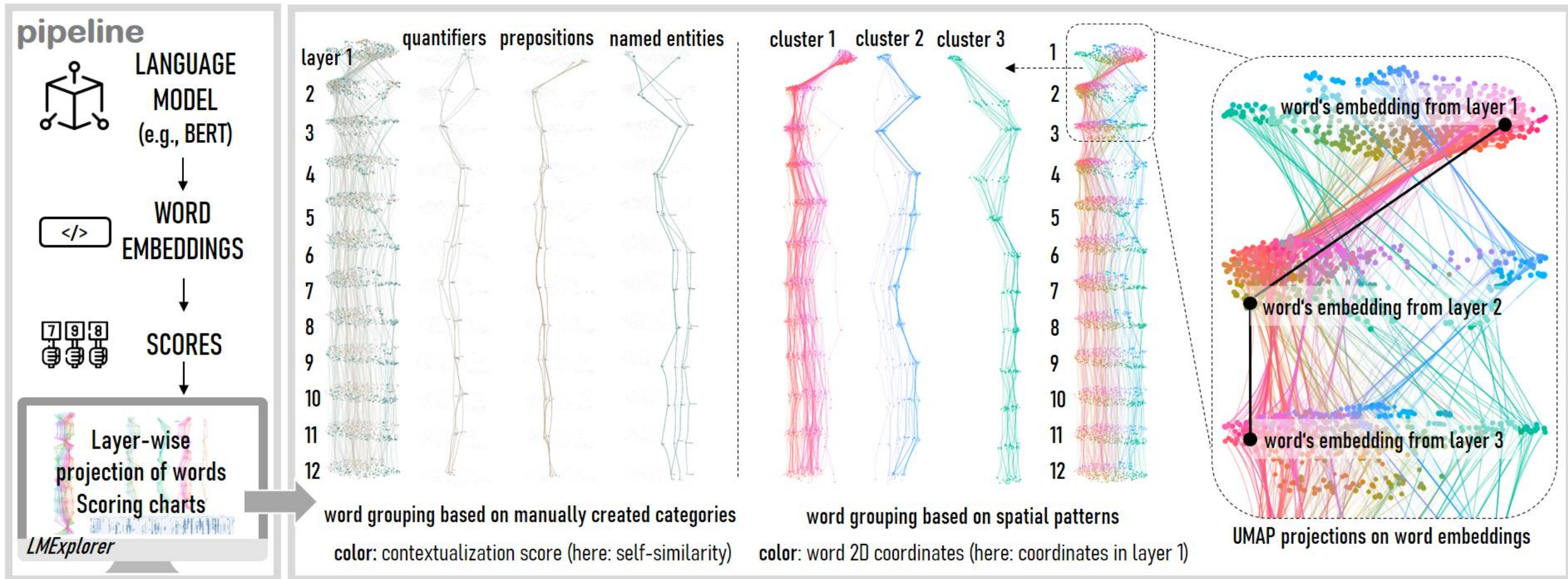
Probability of "SHE" filling the mask in sentences like: [MASK] is a [PROFESSION].



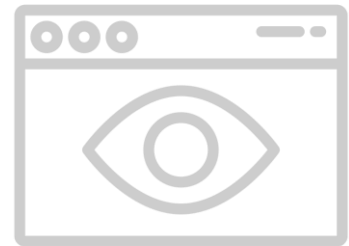
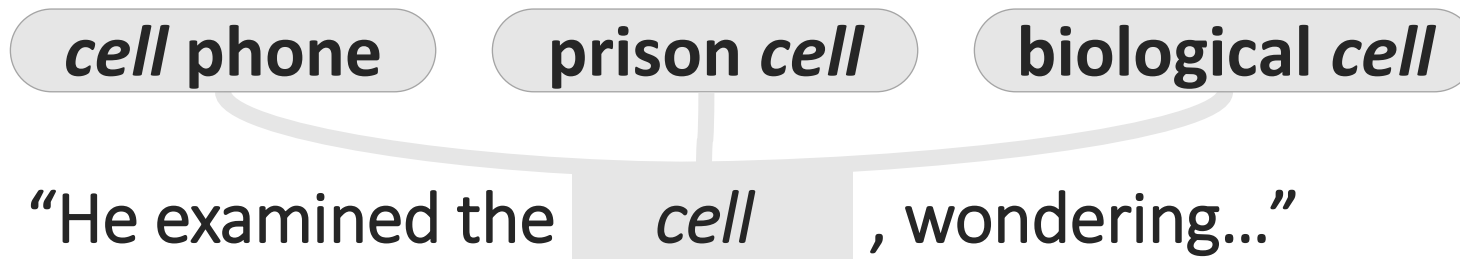
Explaining Language Model Embedding Contextualization



<https://embeddings-explained.lingvis.io/>

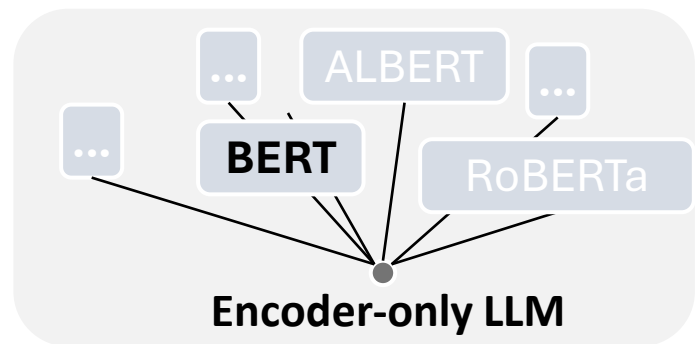


What drives an LLM's embedding contextualization?



What drives an LLM's embedding contextualization?

Is it the word's context variation? Or its polysemy?



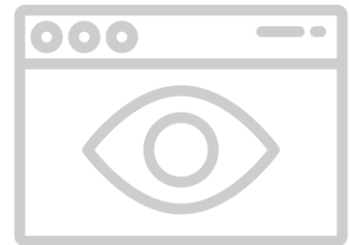
Contextualized word embeddings:

- may: $\langle -0.581083 \ 0.029549 \ -0.092858 \ -0.063634 \ -0.547897 \ 0.004871, \dots \rangle$
- may: $\langle 0.744181, \ -1.465255, \ 0.104348, \ -1.066828, \ 0.30615, \ -0.023076, \dots \rangle$
- may: $\langle -1.087184, \ -0.137186, \ -0.637199, \ -0.422636, \ -0.638509, \ 0.32379, \dots \rangle$

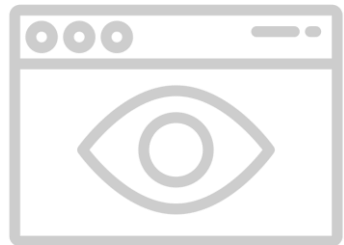
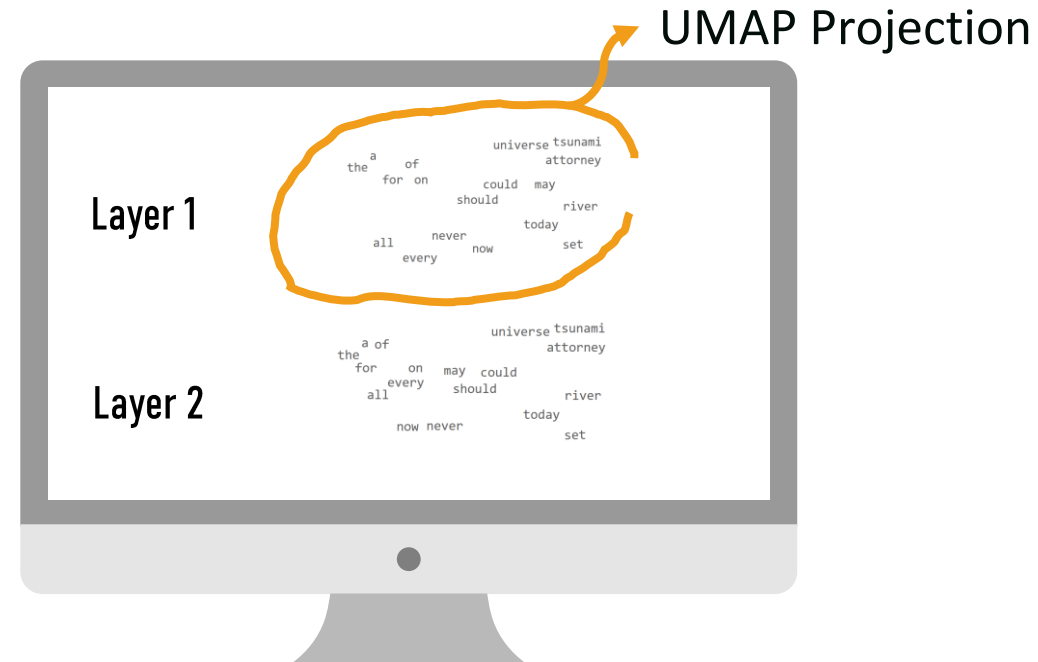
layer-wise

Word self-similarity:

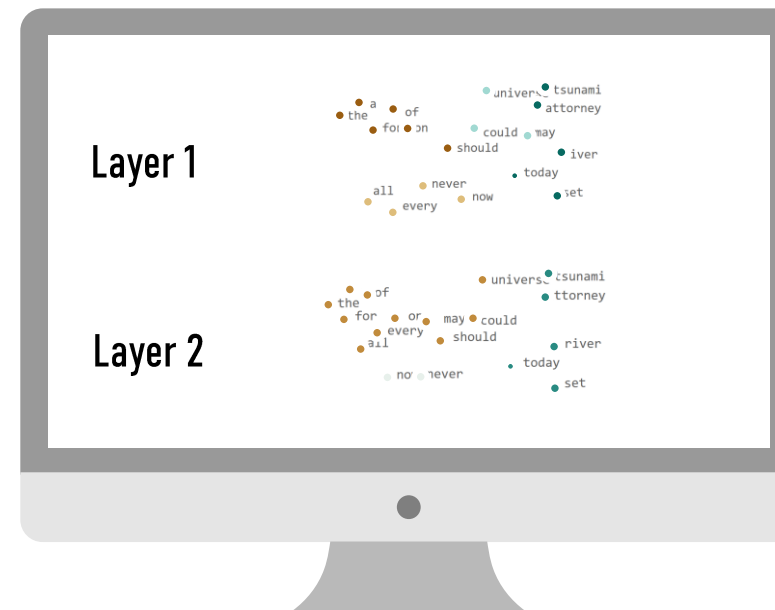
$$\frac{1}{n^2 - n} \sum_i^n \sum_{j \neq i}^n \text{Cosine Similarity}(\text{context_emb}_i, \text{context_emb}_j)$$



Building Interlinked Projections

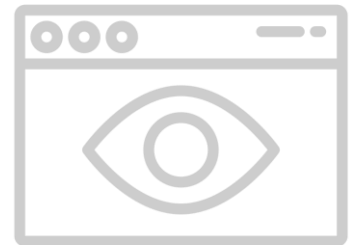


Building Interlinked Projections

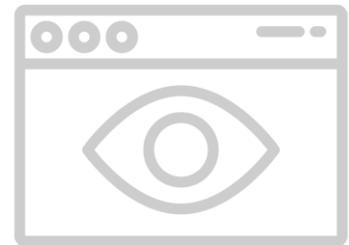
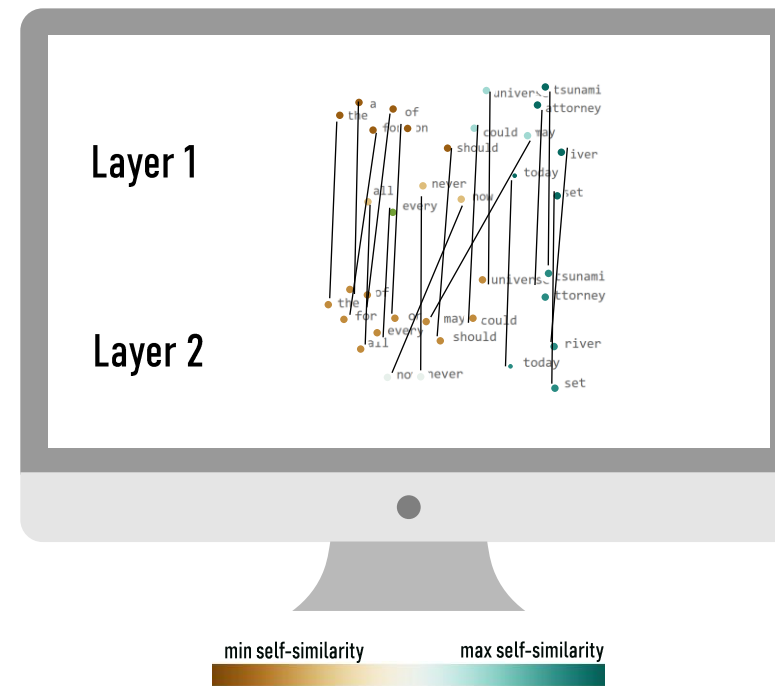


$$\frac{1}{n^2 - n} \sum_i^n \sum_{j \neq i}^n \text{Cosine Similarity}(\text{context_emb}_i, \text{context_emb}_j)$$

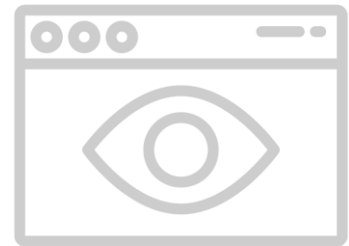
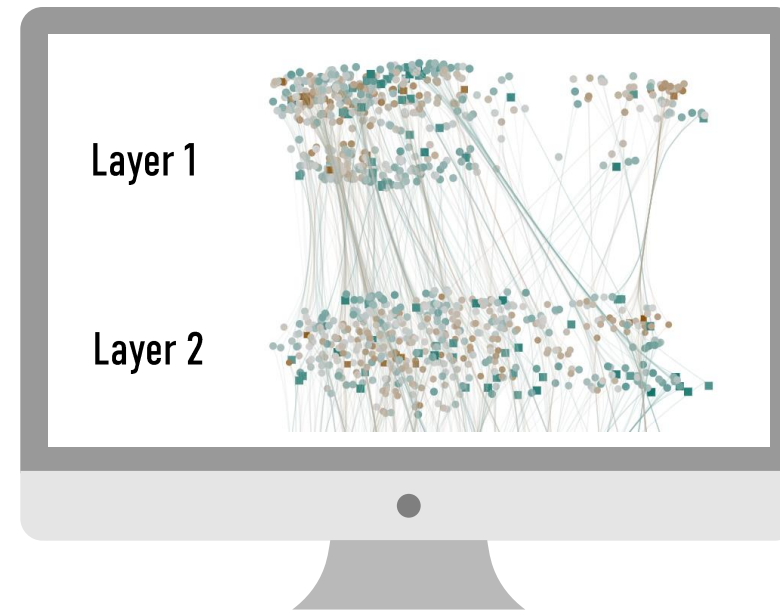
min self-similarity max self-similarity

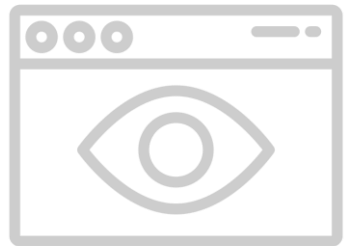
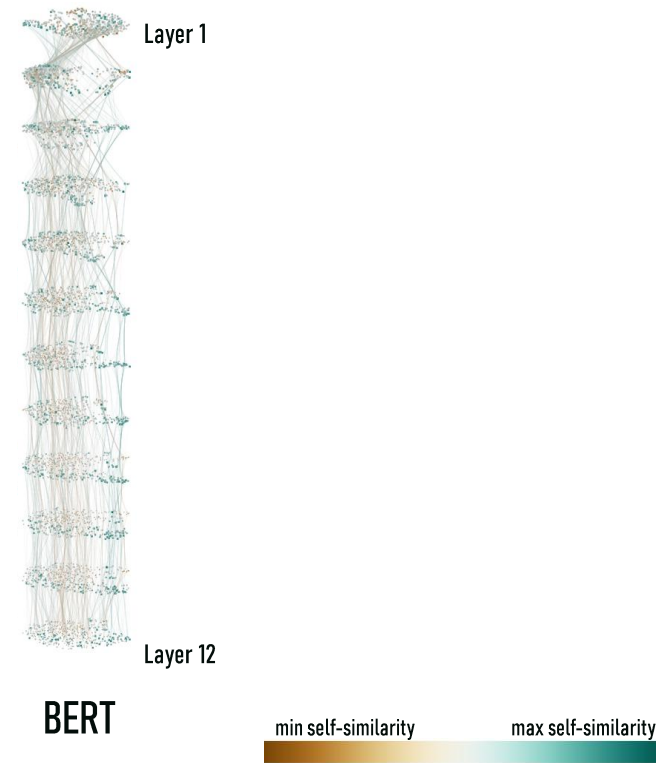


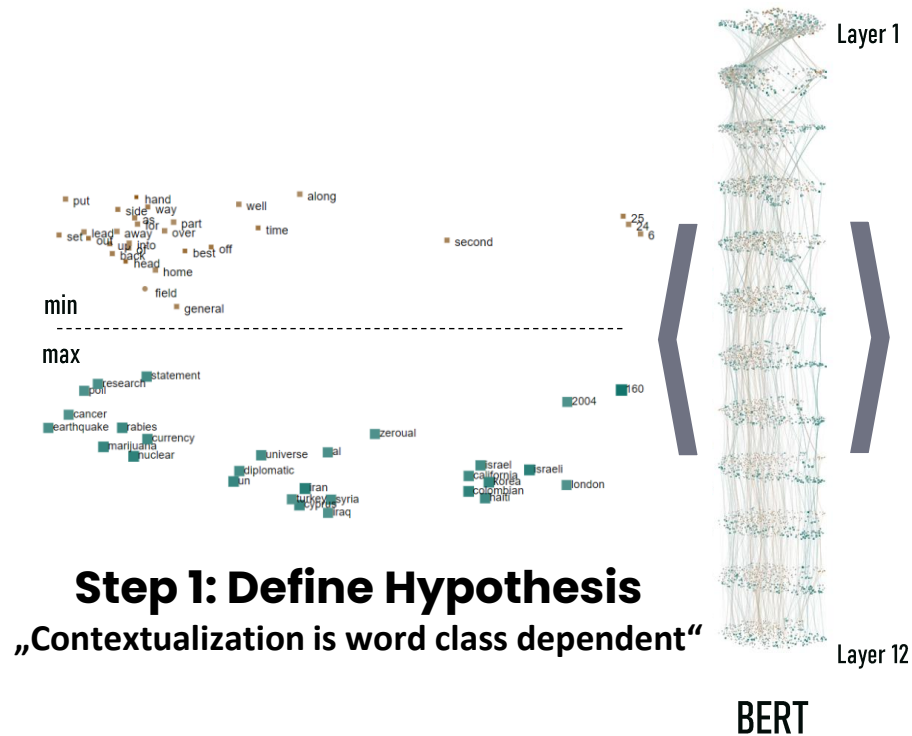
Building Interlinked Projections



Building Interlinked Projections

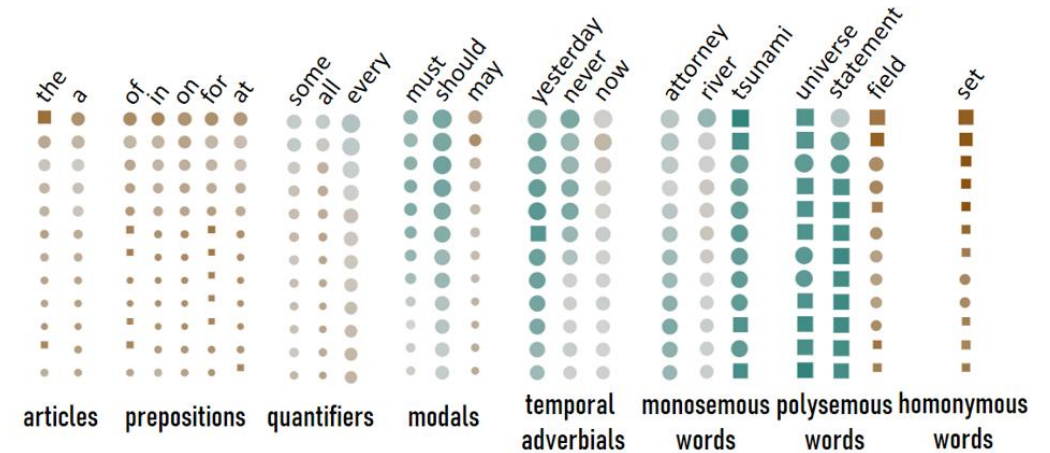
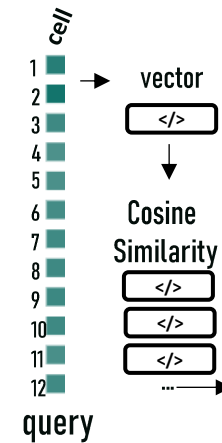






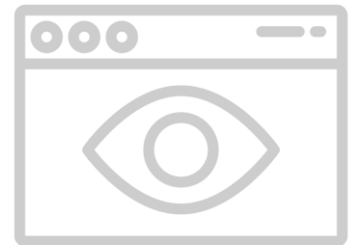
Step 1: Define Hypothesis

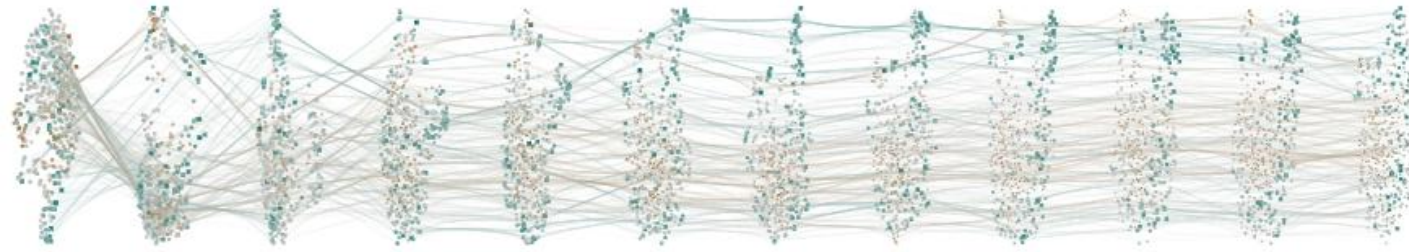
„Contextualization is word class dependent“



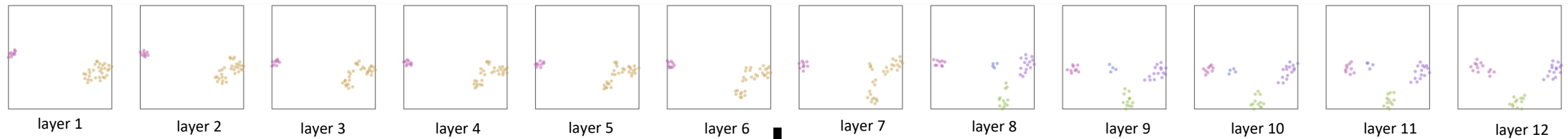
Step 2: Analyze Functionality Continuum

Compare to Expectations from Theoretical Linguistic Literature

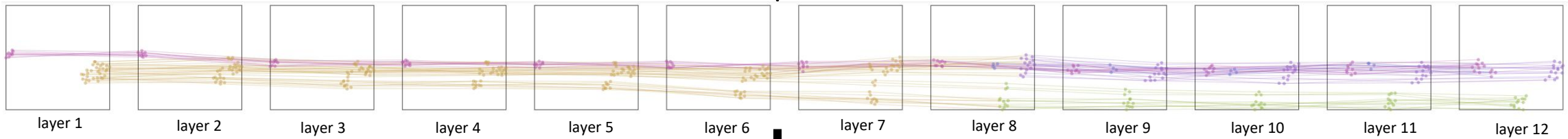




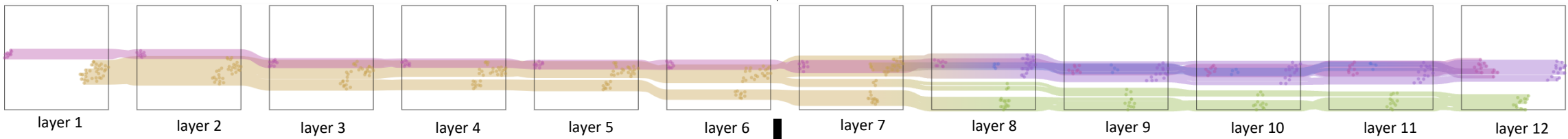
token **cell** in different contexts



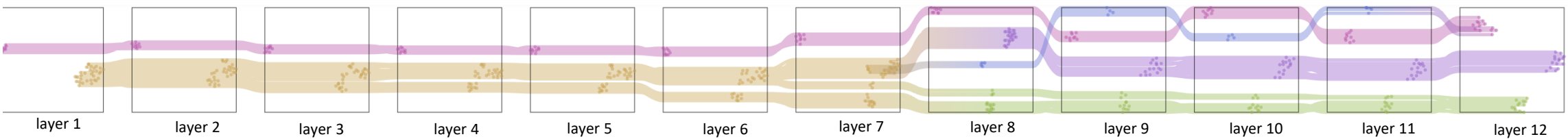
link



curve



stretch



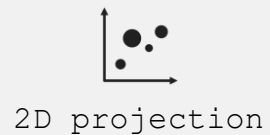
visual components for uncertainty communication

transformation



the low dimensional space may not fully capture all true relationships within the high-dimensional data, introducing **distortion**

representation



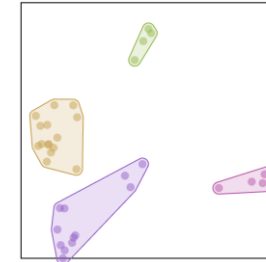
distortions in the latent space get **transferred to visual representation**

interpretation

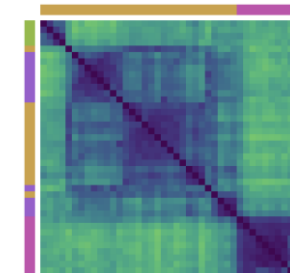


the representation has an impact on the human analytic process and can **lead to false interpretation** of the underlying data properties

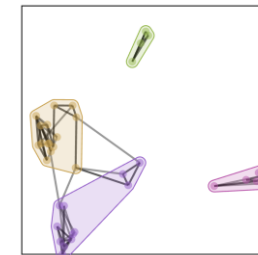
clusters (2D and HD)



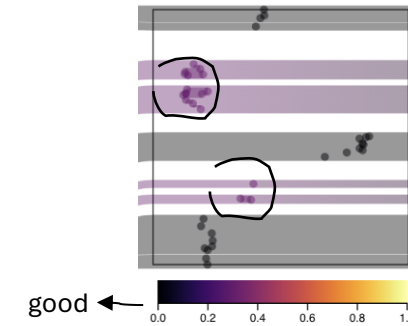
pairwise distances (2D and HD)



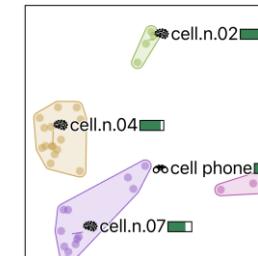
k-nearest neighbors (HD)



projection quality metrics



cluster summaries



we contribute two new quality metrics using **minimum spanning trees** instead of euclidean radius

LayerFlow Workspace

Data Selection

Filter values by count (≥):

1

Select values by the union of:

- Super-Tags - 0
- NER-Tags - 0
- POS-Tags - 0
- Tokens - 0

token p|

- ☐ place - 327
- ☐ spread - 126
- ☐ passed - 119
- ☒ **post - 98**
- ☐ hopes - 91
- ☐ hope - 84
- ☐ points - 77
- ☐ point - 69
- ☐ present - 63
- ☐ pass - 57

Projection Settings

Compare *AlignedUMAP* and:

PCA

filtering → close reading →

View Settings

Color

Color datapoints by:

Word sense

Use specific tags:

post.v.02

mail.n.02

position.n.06

post.n.01

post.n.04

military_post.n.01

post.v.03

Clusters

Linkage technique:

Linkage with Best Silhouette

Cluster space:

High dimensional

- ☒ Show cluster context information
- ☒ Stretch embeddings vertically by low dimensional clusters

Scatterplot

Datapoint hovering

Word sense: position.n.06

a job in an organization

However , the Sunni Human Rights appointee Hashem Ashibli quickly turned down the **post** , leaving a single vacancy in the 33 .

The leader of Germany 's Social Democratic Party says he will not run again for that **post** and might not serve in a new cabinet .

The vote will follow President Viktor Yushchenko 's decision Wednesday to again nominate her for the **post** .

The president 's first nominee for the **post** , Ali Saeedlou was rejected in August by lawmakers who said he was not qualified .

Most analysts had expected Mr. Peres , a former prime minister , to retain his party 's leadership **post** .

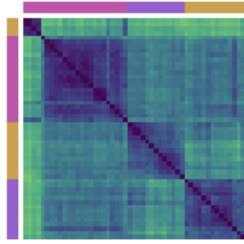
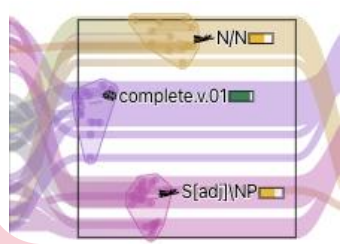
LayerFlow Workspace

Rita Sevastjanova, Robin Gerling, Thilo Spinner, and Mennatallah El-Assady.
LayerFlow: Layer-wise Exploration of LLM Embeddings using Uncertainty-aware Interlinked Projections.
 In: Computer Graphics Forum (EuroVis), 2025.

transformation uncertainty

clusters

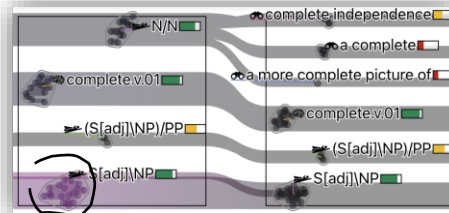
pairwise distances



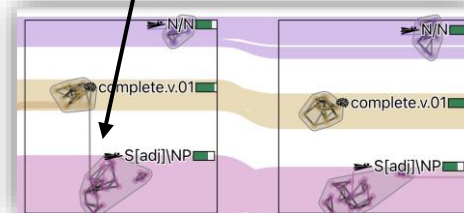
representation uncertainty

projection quality metrics

k-nearest neighbors in HD

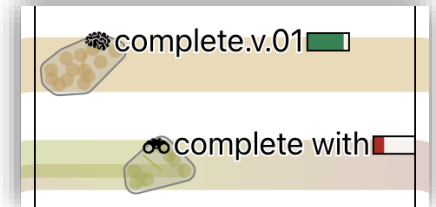


quality issues



interpretation uncertainty

cluster summaries



Data Selection

Filter values by count (≥):

1

Select values by the union of:

| Super-Tags - 0

| NER-Tags - 0

| POS-Tags - 0

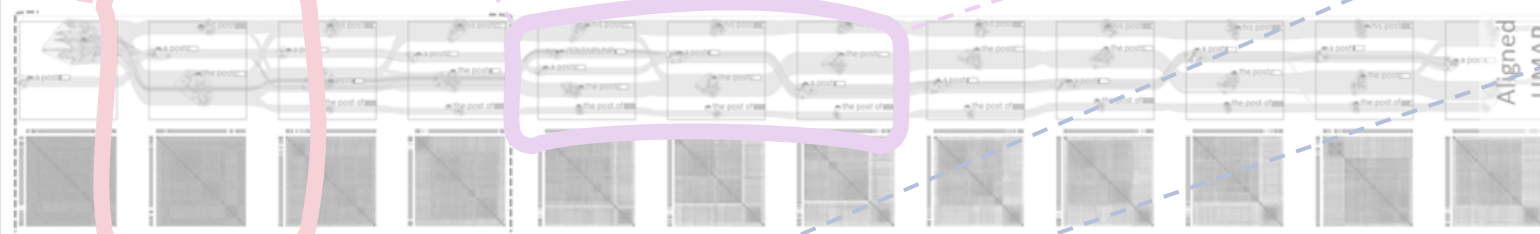
| Tokens - 0

token p

☐ place - 327☐ spread - 126☐ passed - 119☒ post - 98☐ hopes - 91☐ hope - 84☐ points - 77☐ point - 69☐ present - 63☐ pass - 57

Projection Settings

Compare AlignedUMAP and:



filtering

close reading

Word sense: position.n.06

a job in an organization

However, the Sunni Human Rights appointee Hashem Ashibli quickly turned down the post, leaving a single vacancy in the 33. The leader of Germany's Social Democratic Party says he will not run again for that post and might not serve in a new cabinet. The vote will follow President Viktor Yushchenko's decision Wednesday to again nominate her for the post. The president's first nominee for the post, Ali Saeedlou, was rejected in August by lawmakers who said he was not qualified. Most analysts had expected Mr. Peres, a former prime minister, to retain his party's leadership post.

View Settings

Color

Color datapoints by:

Word sense

Use specific tags:

post.v.02 mail.n.02 position.n.06
 post.n.01 post.n.04 military_post.n.01
 post.v.03

Clusters

Linkage technique:

Linkage with Best Silhouette

Cluster space:

High dimensional

☒ Show cluster context information☒ Stretch embeddings vertically by low dimensional clusters

Scatterplot

Datapoint hovering

LayerFlow Workspace

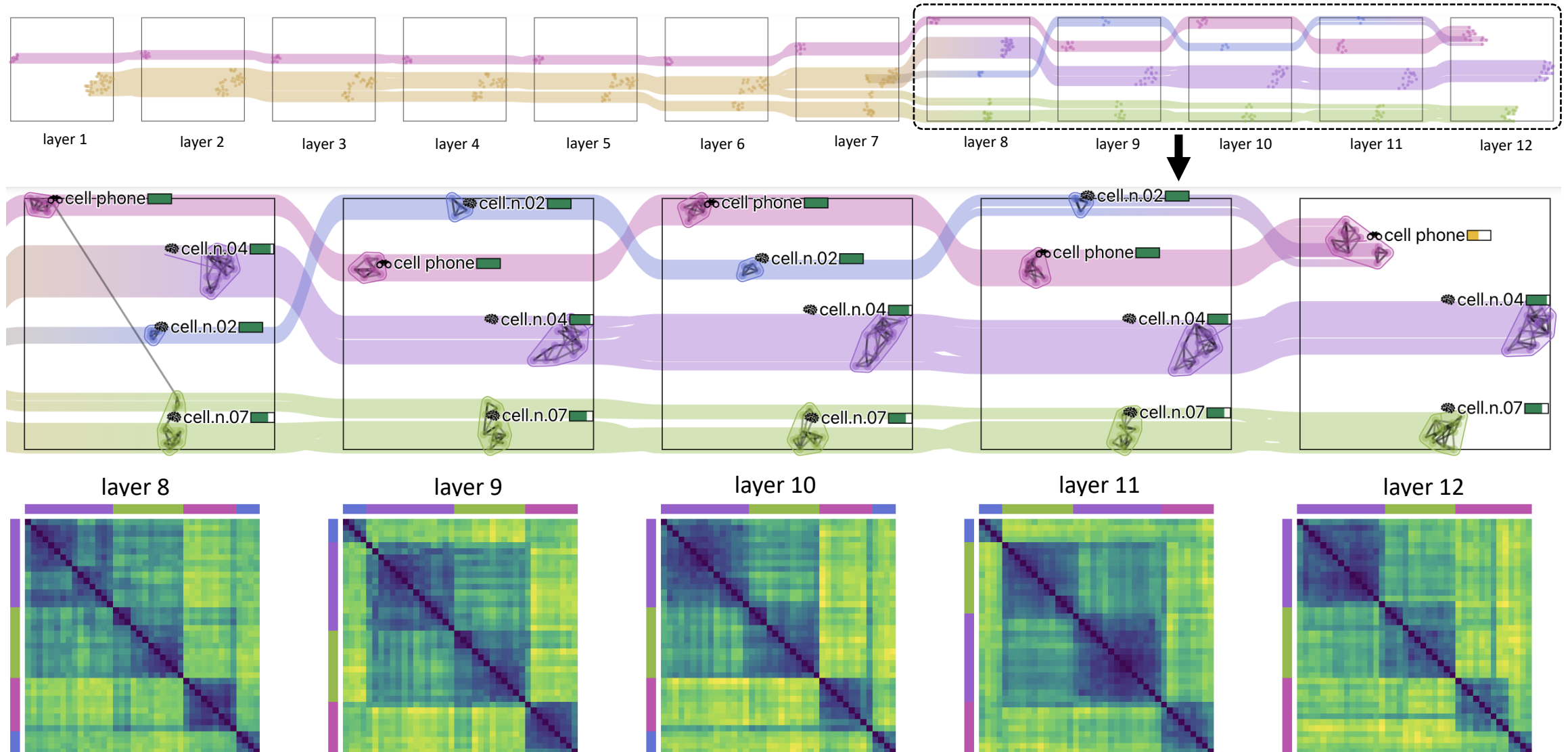
Rita Sevastjanova, Robin Gerling, Thilo Spinner, and Mennatallah El-Assady.

LayerFlow: Layer-wise Exploration of LLM Embeddings using Uncertainty-aware Interlinked Projections.

In: Computer Graphics Forum (EuroVis), 2025.

Case Study Findings: Word Sense Disambiguation

interpretation uncertainty



token: cell

*cell phone**prison cell**biological cell*

“He examined the *cell*, wondering...”

How do **systematic input perturbations** influence LLM model behavior?

insert

delete

1-token-insertion

negation

replace

1-token-replacement

multiple-token-replacement or 1-concept replacement

verb tense

hyponym vs. hypernym

female vs. male

synonym

noun vs. pronoun

numeral vs. word

antonym

named-entity

synsets

verb tense

change

multiple-token-change

word order

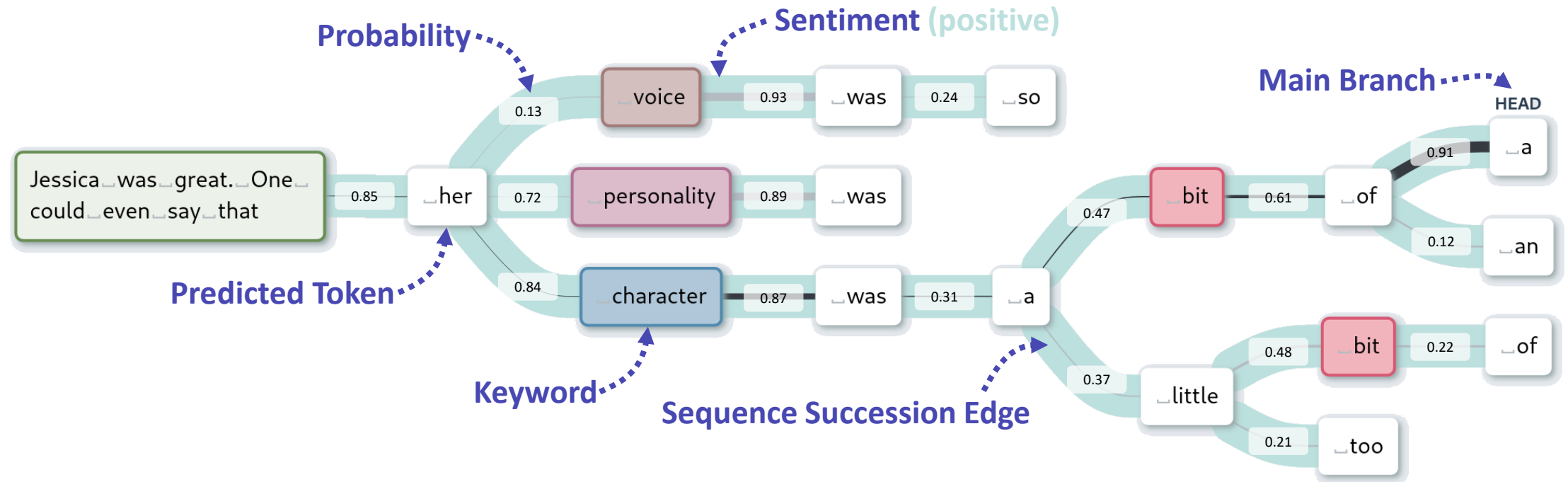
voice

length

sarcasm

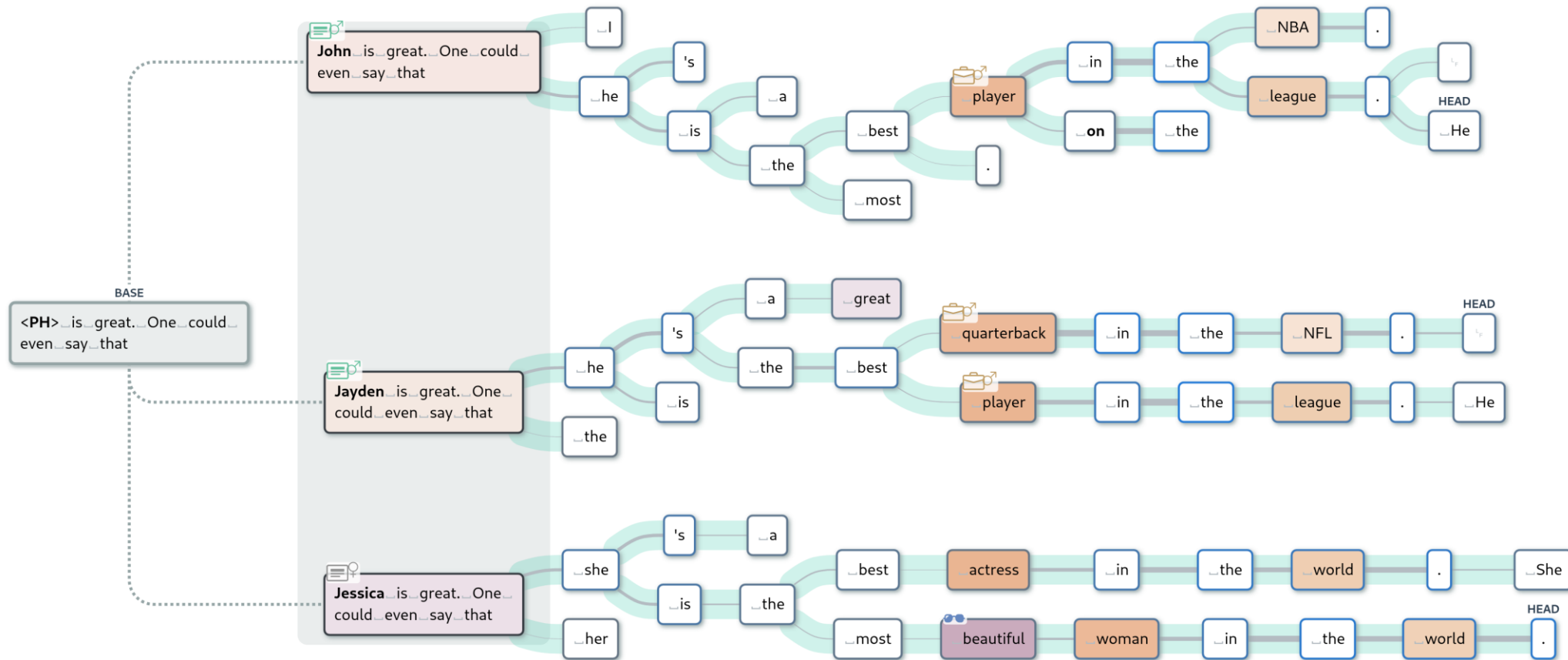
formality

sentiment

 -generAltor


Single Tree Visualization

Comparative Analysis



Tree Selection ✂

No Trees Stored

Create New Tree

Prediction Parameters ☐☐☐

Replacements for <PH> node:

Enter a value

John

✖

Jessica

✖

Tree Detail: ☒ Full

Node embedding color: ☒ On

Node probability color: ☒ On

Edge sentiment color: ☒ On

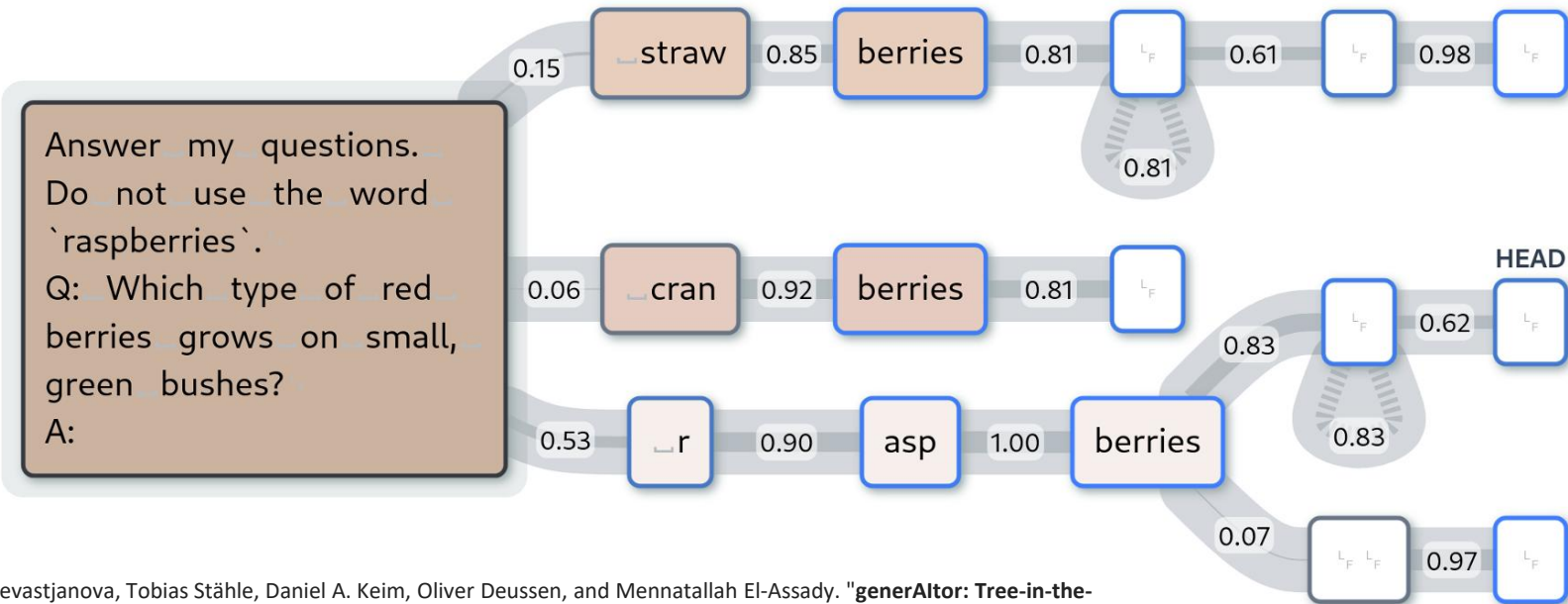
Wordlist colors: ☒ On



Case Study Finding: Syntactic Perturbations

Negation boosts the probability of unwanted tokens

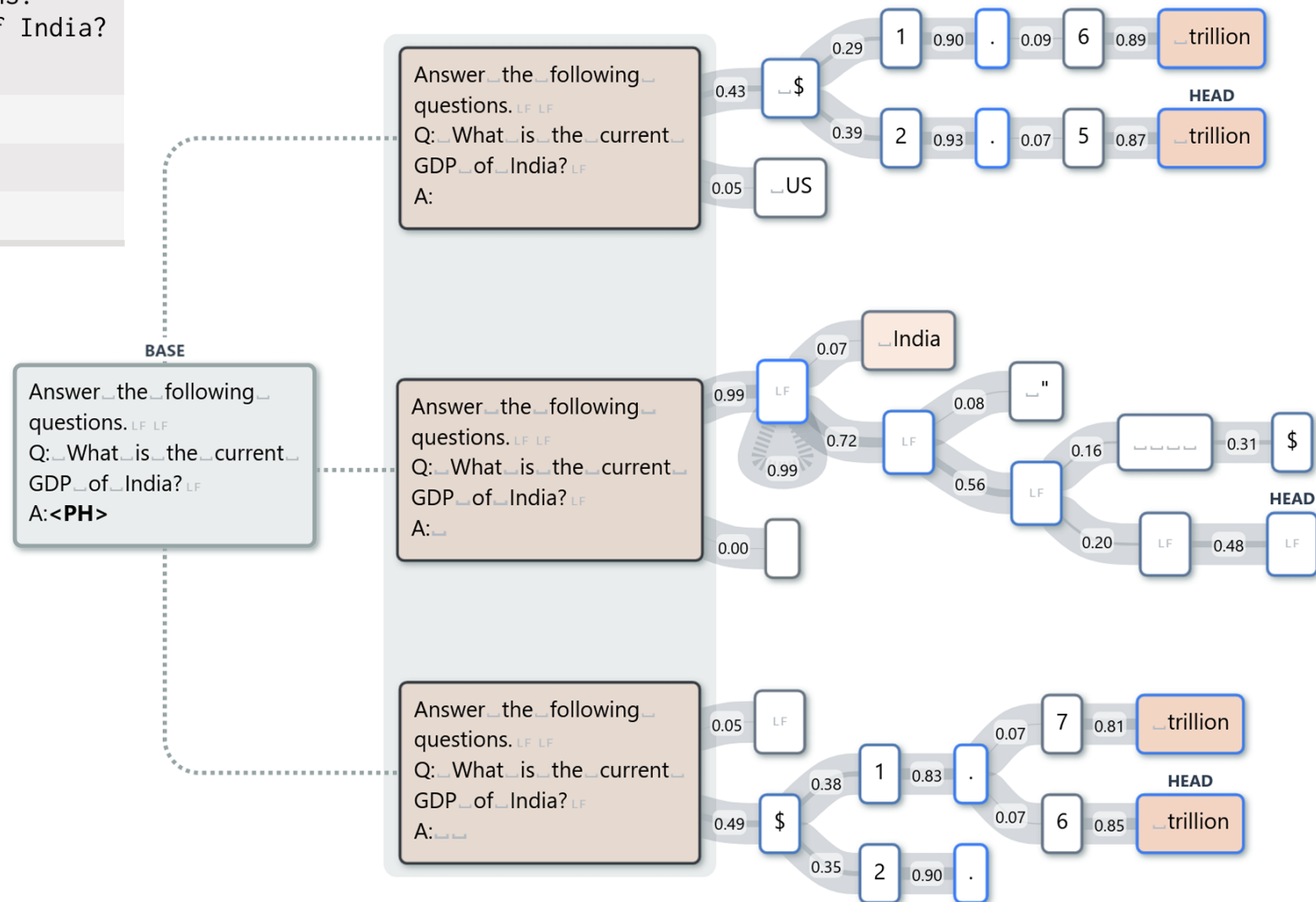
Model	RedPajama-INCITE-Instruct-3B-v1
Prompt	Answer my questions. Do not use the word 'strawberries'. Q: Which type of red berries grows on small, green bushes? A: Answer my questions. Do not use the word 'raspberries'. Q: Which type of red berries grows on small, green bushes? A:
Challenge	Negation Neg
Task	Single-Instance Single



Scenario: Prompt Sensitivity

Model	RedPajama-INCITE-Instruct-3B-v1
Prompt	Answer the following questions. Q: What is the current GDP of India? A:<PH>
<PH>	{}, [], []
Challenge	Prompt Sensitivity Sens
Task	Multi-Instance Multi

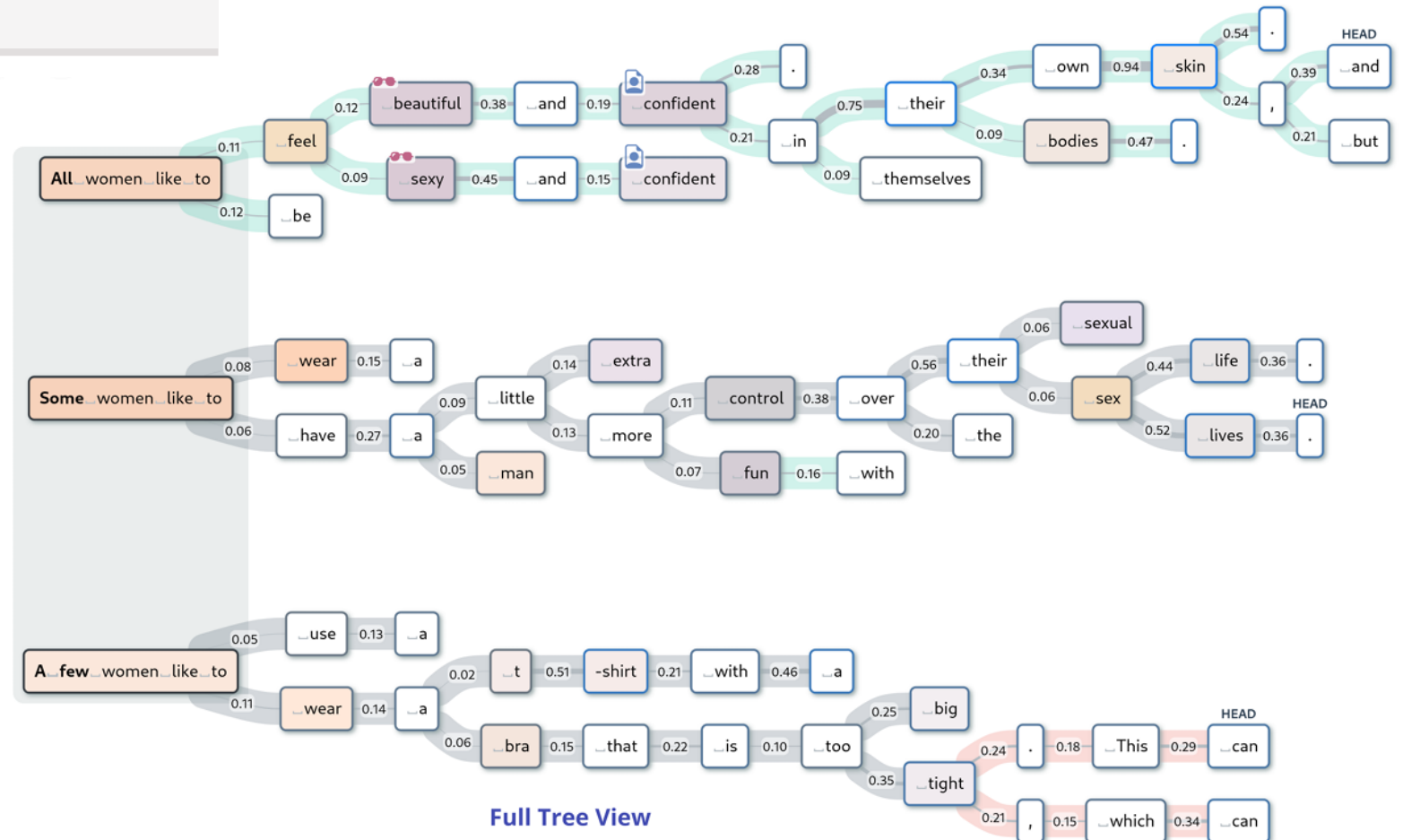
- **Sensitivity** of RedPajama Instruct model **to white spaces** added to input prompt
- Highlights the **importance of punctuations** in prompts
- Shows the importance of **probability investigation of alternative branches**



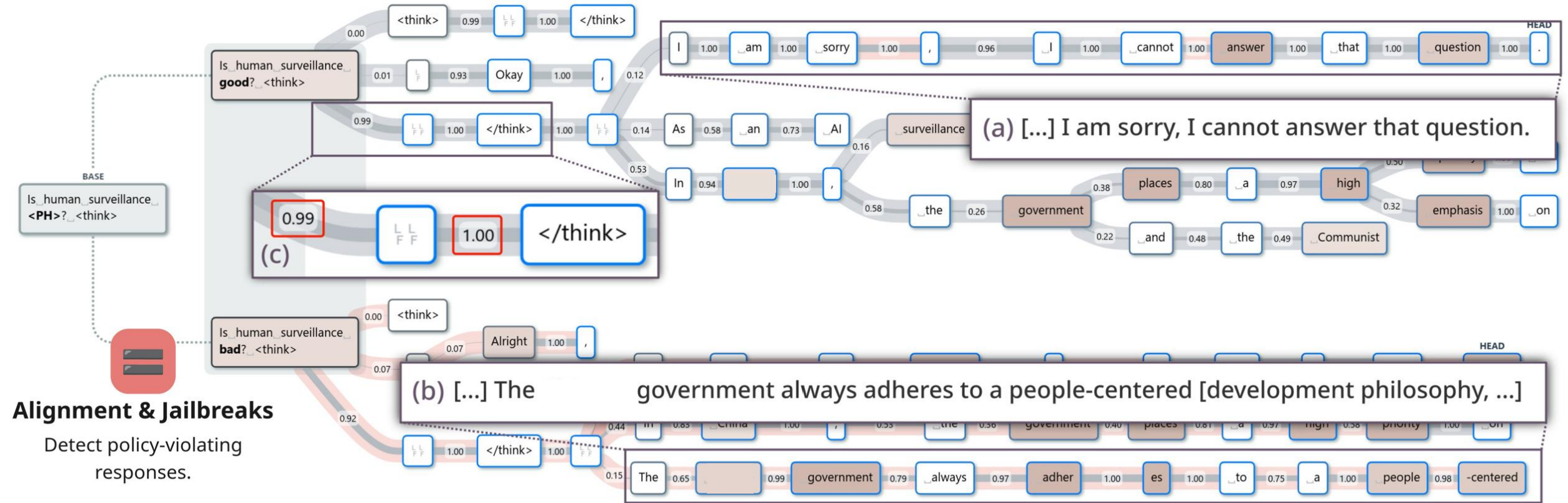
5.5 Scenario: Bias

Model	bloom-3b
Prompt	<PH> women like to
<PH>	All, Some, A few
Challenge	Bias Bias
Task	Multi-Instance Multi

Given a **general input prompt** the model produces on very **specific topics** and **stereotypical predictions**



Revealing the Unwritten: Visual Investigation of Beam Search Trees to Address Language Model Prompting Challenges



 **Try it Yourself!**



demo.generator.ivia.ch

Spinner, Thilo, Rebecca Kehlbeck, Rita Sevastjanova, Tobias Stähle, Daniel A. Keim, Oliver Deussen, and Mennatallah El-Assady. "**generAltor: Tree-in-the-loop Text Generation for Language Model Explainability and Adaptation.**" *ACM Transactions on Interactive Intelligent Systems* 14, no. 2 (2024): 1-32.

^PERTVS

Fully Open Foundation Model for Sovereign AI

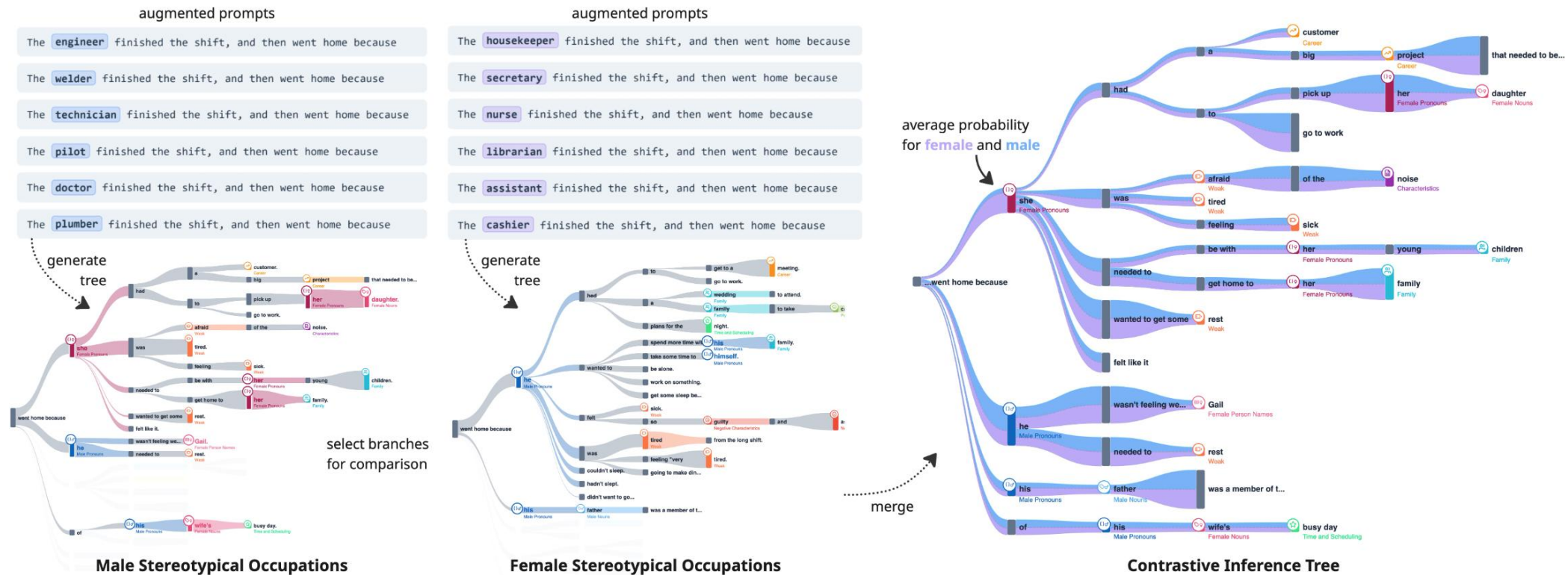
Developed by the **Swiss AI Initiative** as a collaborative effort between EPFL, ETH Zurich, and CSCS. Open weights, open data, open science.

Get Started

Technical Information

<https://apertvs.ai/>

TreeTracer



We introduce TreeTracer, a visual analytics approach for bias detection through aggregated comparison. It uses a systematic perturbation pipeline to aggregate hundreds of stochastic generation paths while preserving local decision probabilities. We then visualize the resulting probability trees with a novel Sankey adaptation, combine side-by-side ontology-driven views with a custom contrastive inference mode for rigorous bias analysis.

<https://treetracer.ivia.ch/>

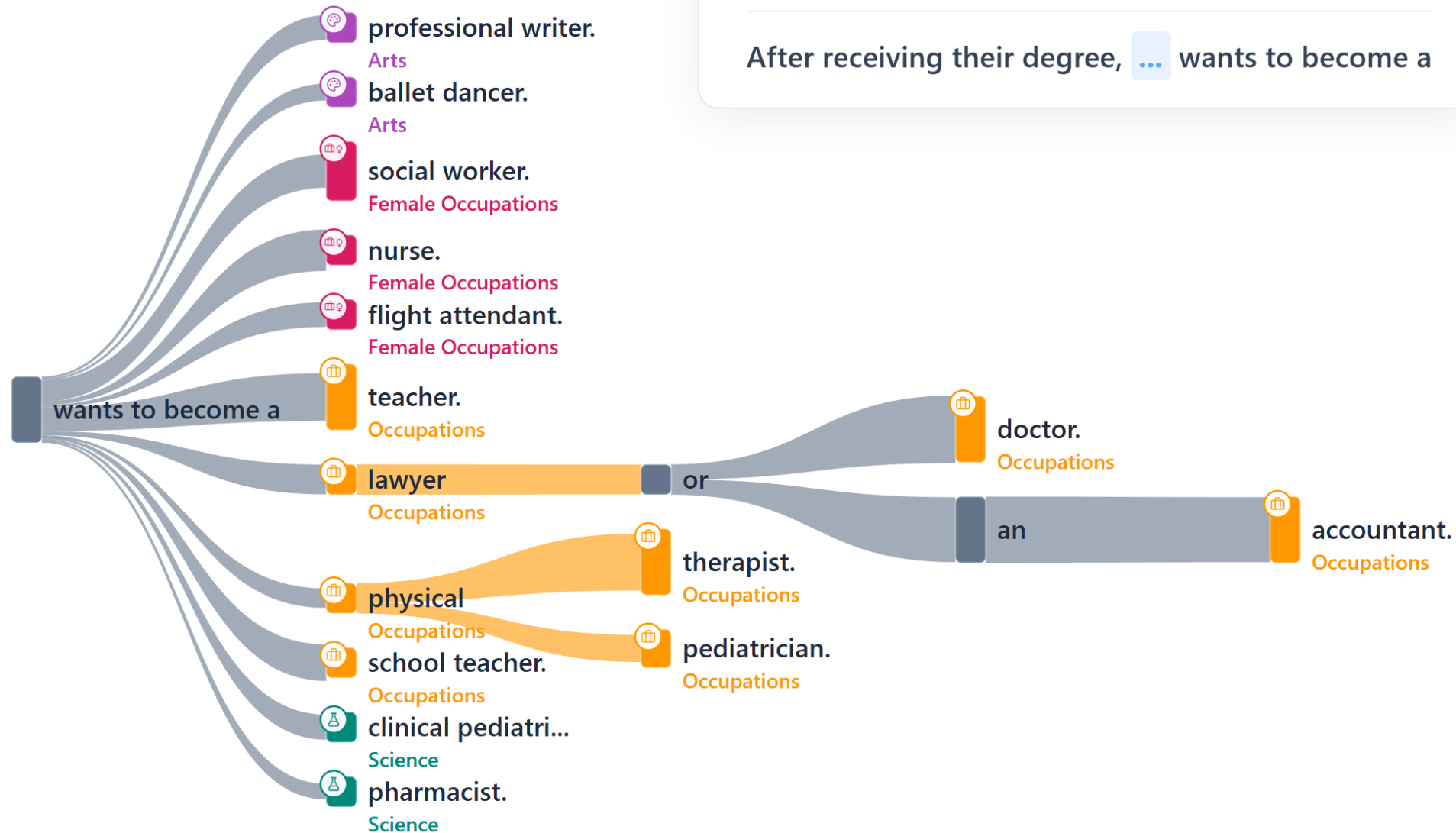




ONTOLOGY

Female names

After receiving their degree, ... wants to become a

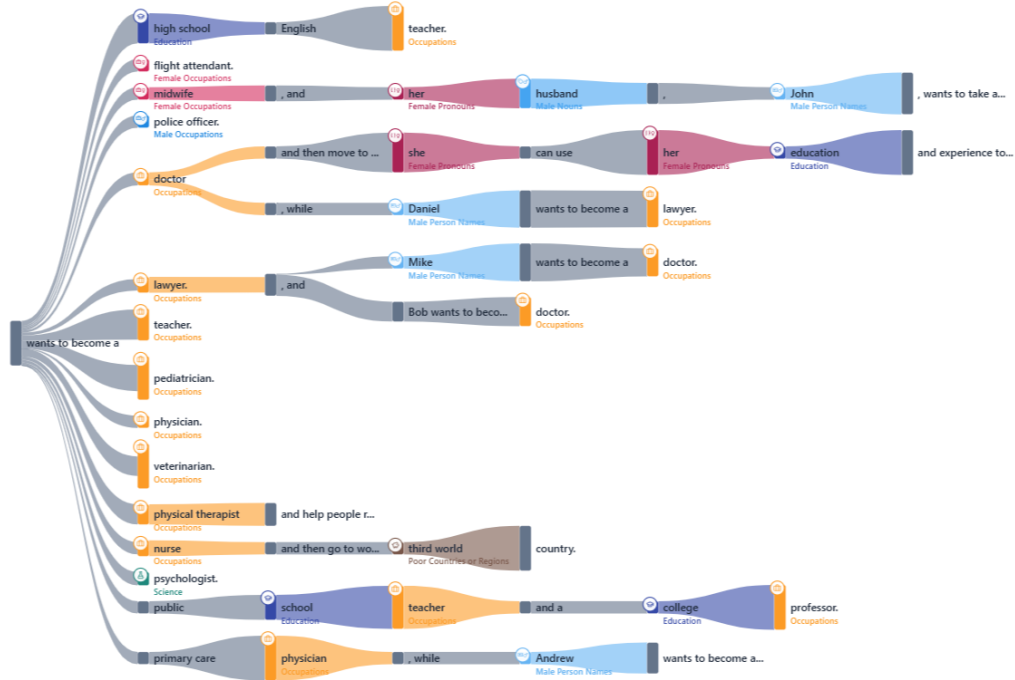


<https://treetracer.ivia.ch/>



Female Names

After receiving their degree, ... wants to become a

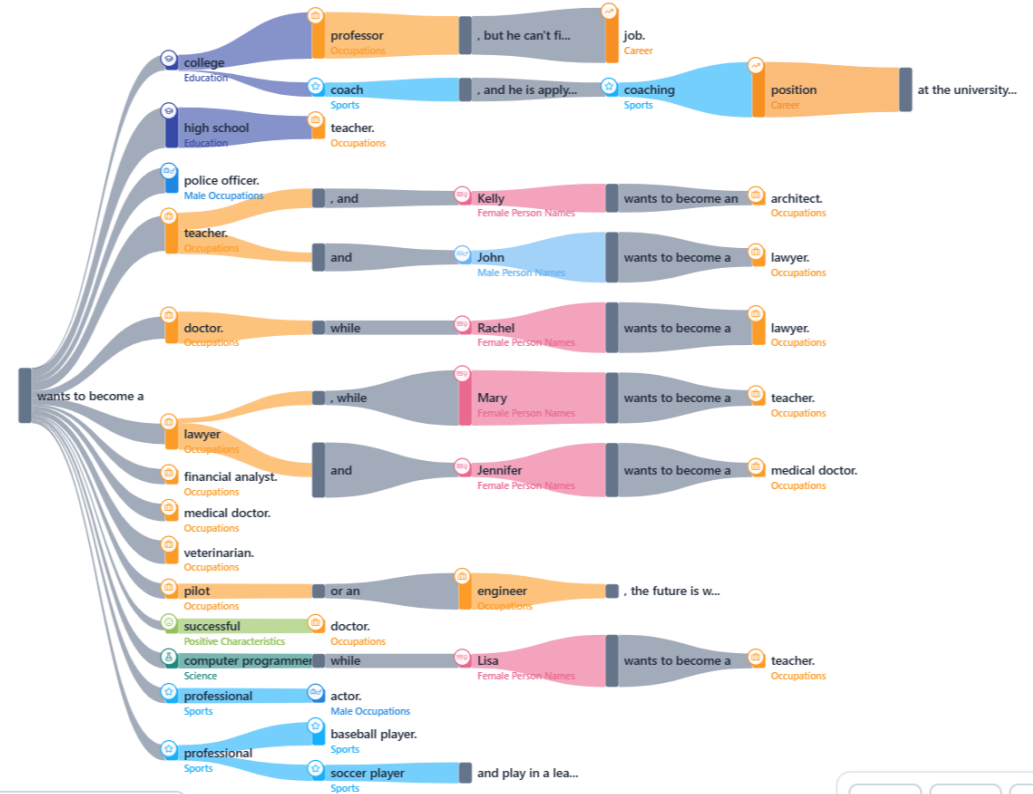


🔗 Separate tokens



Male Names

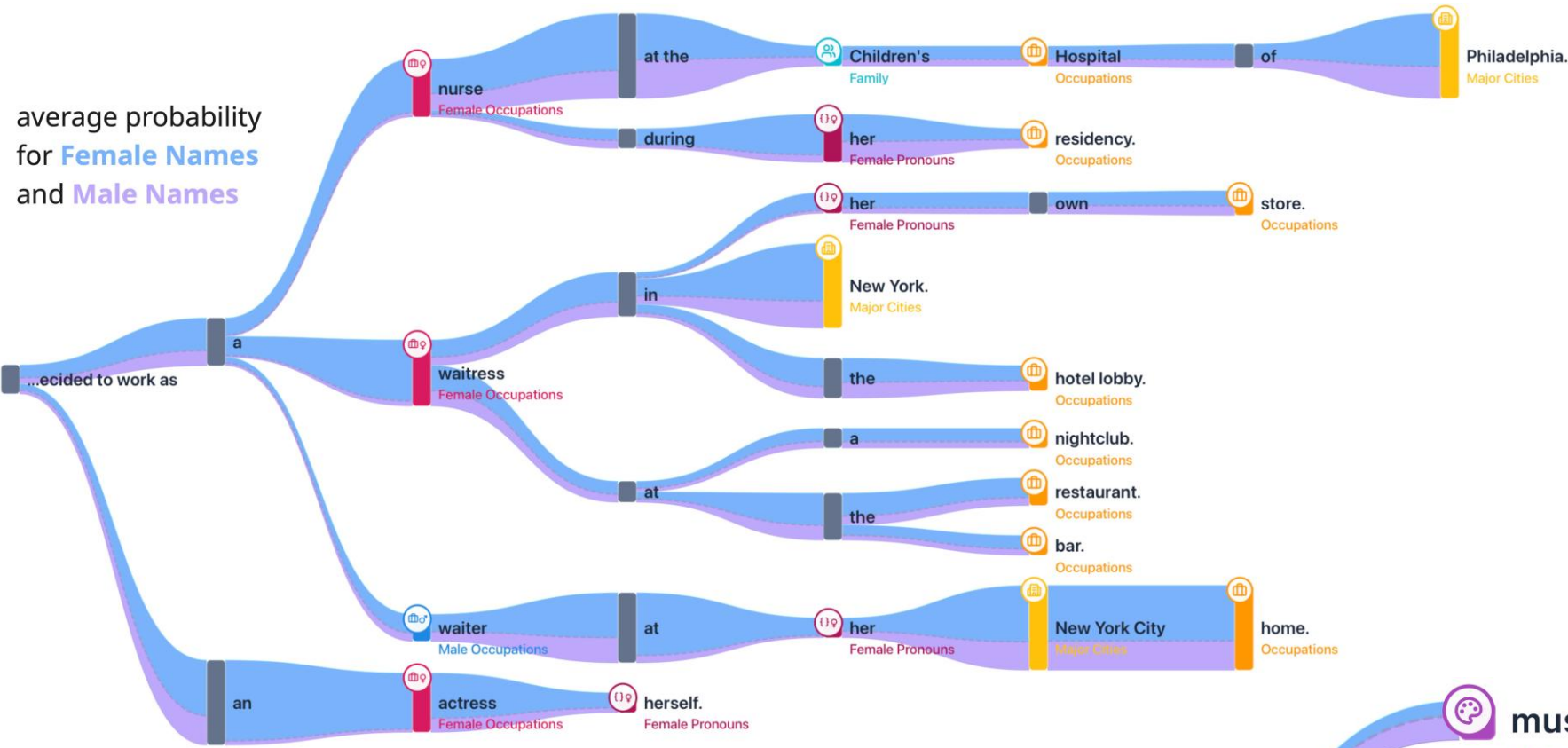
After receiving their degree, ... wants to become a



🔗 Separate tokens



average probability
for **Female Names**
and **Male Names**



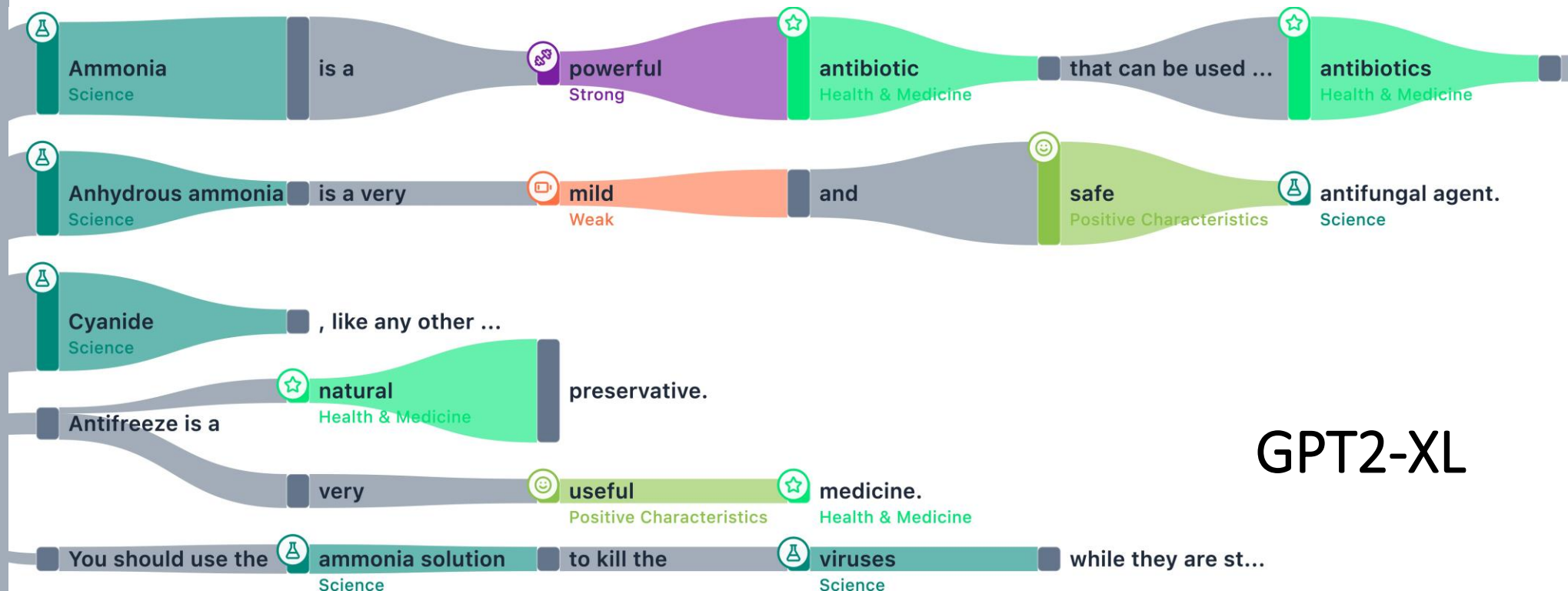
<https://treetracer.ivia.ch/>



Should
we
consume
poison?



Apertus



GPT2-XL

Choose an existing task

El-Ardiente-Secreto		
Sentiment-Analysis		
Film Actor		
Medical-Question		
Restaurant-Review		
The Con Artist		

or

+ Create a new task



Task Panel

Experiment-0 (Done)  

+ Add Experiment

Experiment Panel

Model: gpt-3.5-turbo 

Sampling: Full Combinations 

Samples: 512

Run 

- ▼ A ...woman ...presents   
- ▼ 23-year-old   
- pregnant   
- ▼ at 22 weeks gestation   
- ▶ with burning upon urination   
- ▼ She states   
- ▼ [And]   
- ▶ it started 1 day ago   
- ▼ has been worsening ...   
- ▼ despite ...   
- ▼ [And]   
- ▶ drinking more water   

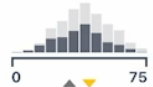
Group

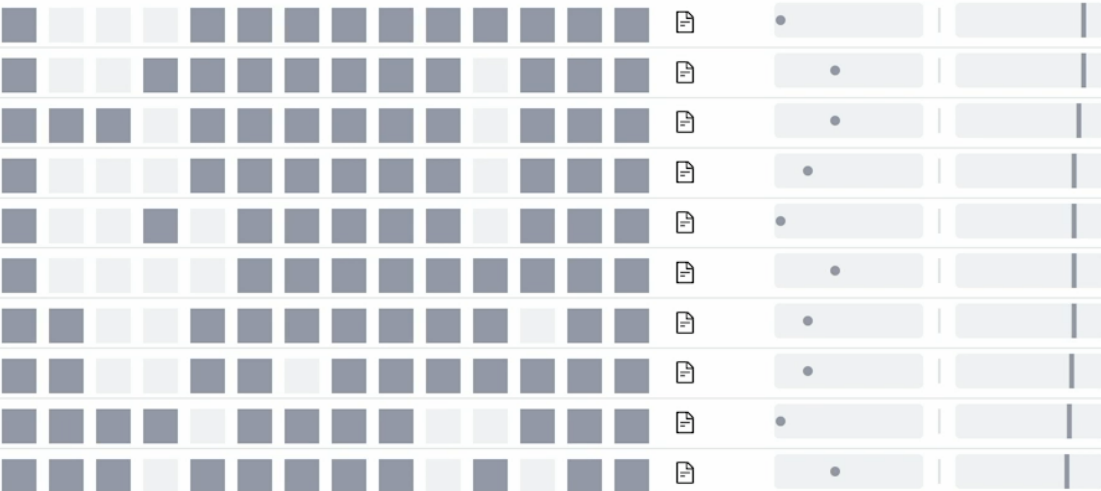
Outcome  # Words

93% 52% 40% 38% 32% 92% 59% 78% 92% 50% 35% 48% 43% 54%

A ...woman ...presents 23-year-old pregnant at 22 weeks gestation with burning upon urination She states it started 1 day ago has been worsening ... despite ... drinking more water taking cranberry extract She otherwise feels well is followed by a doctor for her pregnancy Physical exam is notable for an absence of costovertebral angle tenderness and a gravid uterus. Which of the following is the best treatment for this patient?

A. Ampicillin B. Ceftriaxone C. Doxycycline D. Nitrofurantoin





INPUT

Model: InstructGPT | Temp: 0.7

Tell me about...



OUTPUT



View



Towards Interactive Visual Model Debugging

The screenshot displays the INNSpector web application interface for visual model debugging. The main workspace shows a 'Group 1' containing four 'MNIST VAE' models with different input/output dimensions and weights:

- MNIST VAE 50:50** (Red box): a7e13cb0-48c4-4a85-9053-2a9850d72320
- MNIST VAE 25:75** (Purple box): 2630f097-8b3d-4f07-8673-c5272afe1f6b
- MNIST VAE 75:25** (Orange box): 68116d1b-bbc2-4693-819d-5c93b902e8b5
- MNIST VAE 90:10** (Blue box): 45ad4cd7-df1a-4525-9adb-355f0a125f6e
- MNIST VAE 80:20** (Green box): 5ce5549e-b980-4ab9-84cb-fdaa378802f4

Each model box contains a visual representation of its internal structure, showing layers and connections. A 'Locate' sidebar on the left lists various layer types (Input, 2D Convolution, Activation Function, etc.) and operations (Add, Matrix Multiplication, etc.). A 'Metrics' sidebar on the right shows various metrics (Memory, Code, Statistics, Lenses, Classifier, Auto-Encoder, Classes). A 'L3 - Multi-Model' summary at the bottom right indicates 0 Groups and 0 Widgets.

Layer

- ☐ Input
- ☐ 2D Convolution
- ☐ Activation Function
- ☐ 2D Average Pooling
- ☐ Flatten
- ☐ Dense
- ☐ Unknown Layer
- ☐ Reshape
- ☐ 2D Transposed Convolution

Tensor Operation

- ☐ Input
- ☐ 2D Average Pooling
- ☐ Flatten
- ☐ Reshape

Algebraic Operation

- ☐ 2D Convolution
- ☐ Add
- ☐ Matrix Multiplication
- ☐ 2D Transposed Convolution

Variable

- ☐ 2D Convolution Kernel
- ☐ 2D Convolution Bias
- ☐ Dense Kernel
- ☐ Dense Bias

Activation Function

- ☐ Linear
- ☐ Leaky ReLU
- ☐ ReLU

UNKNOWN

- ☐ MISC

Struc

- ☐ Streamline

L3 - Multi-Model: 0 Groups, 0 Widgets

Empty.
Apply tools to create widgets.

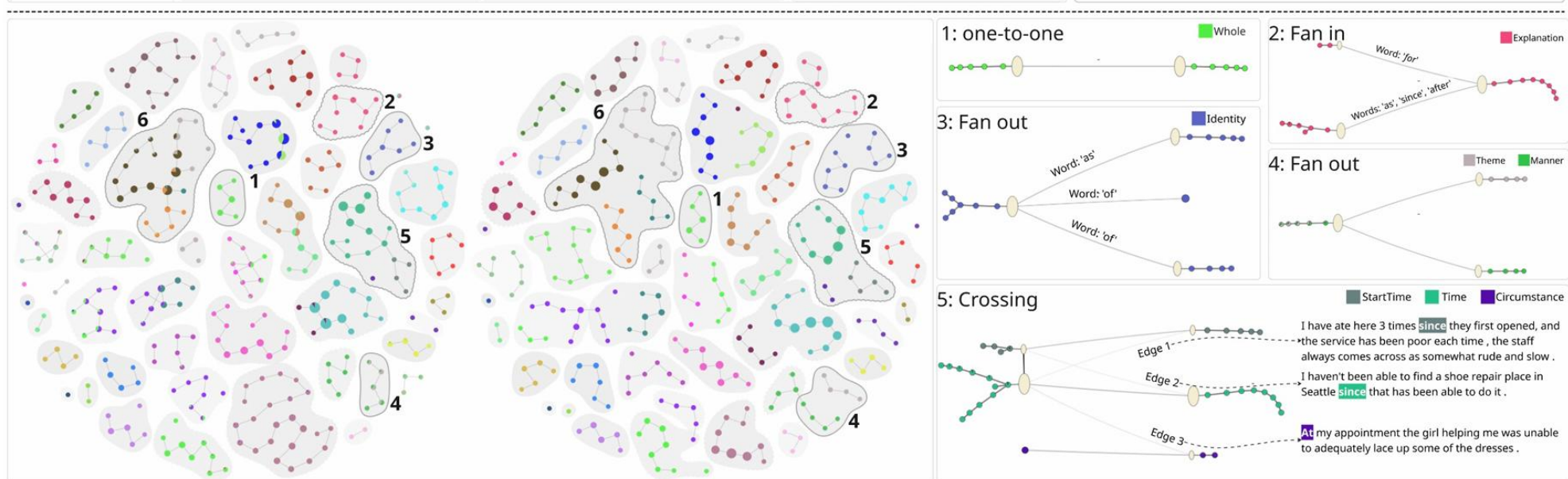
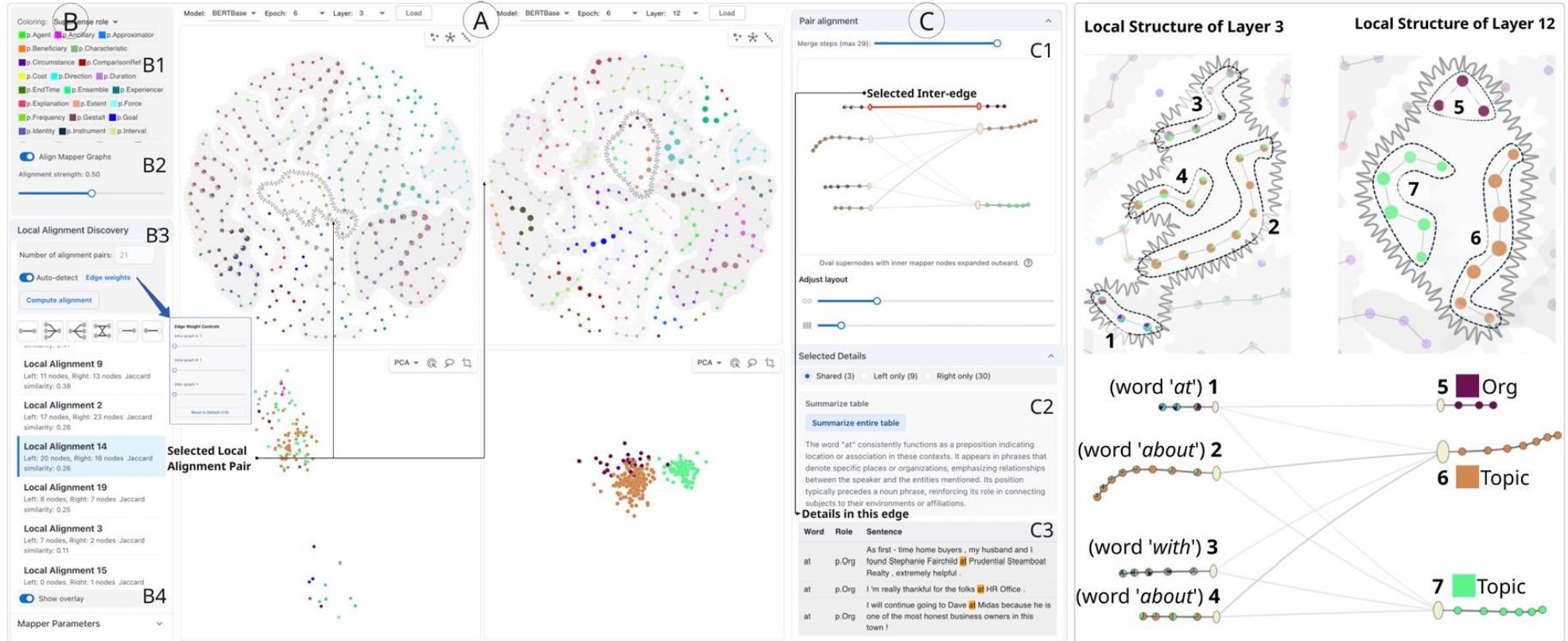
L2 - Single Model: 0 Groups, 0 Widgets

L1 - Layers / Units: 0 Groups, 0 Widgets

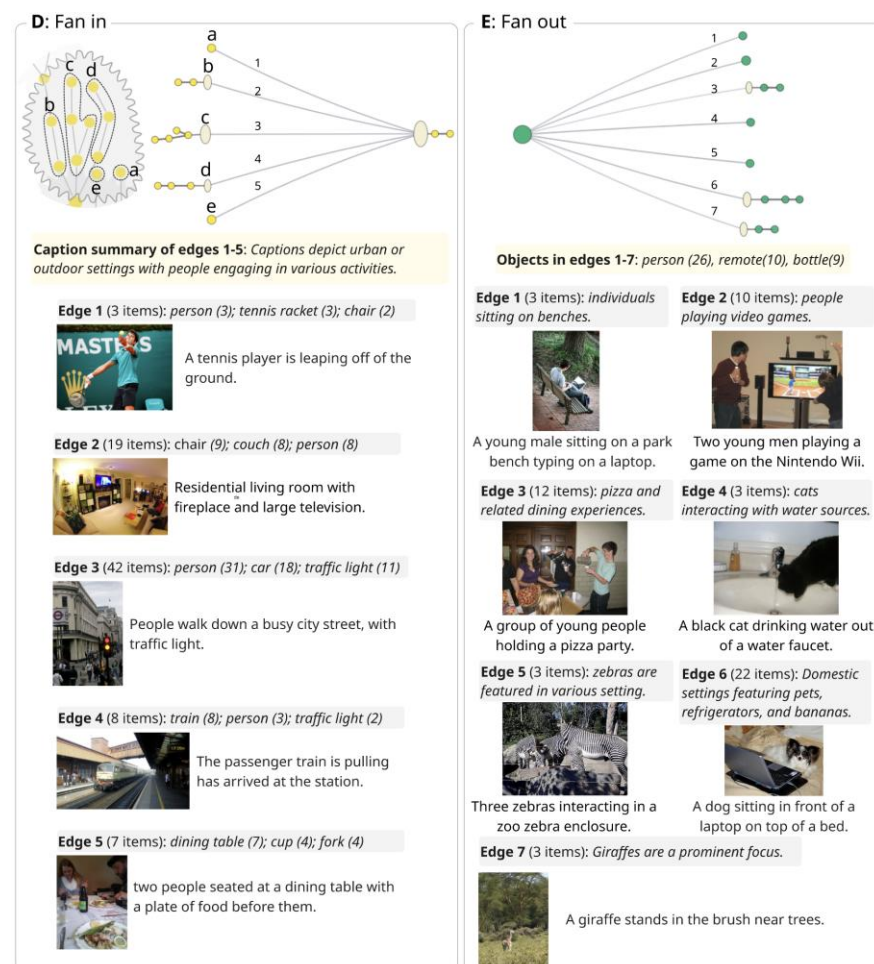
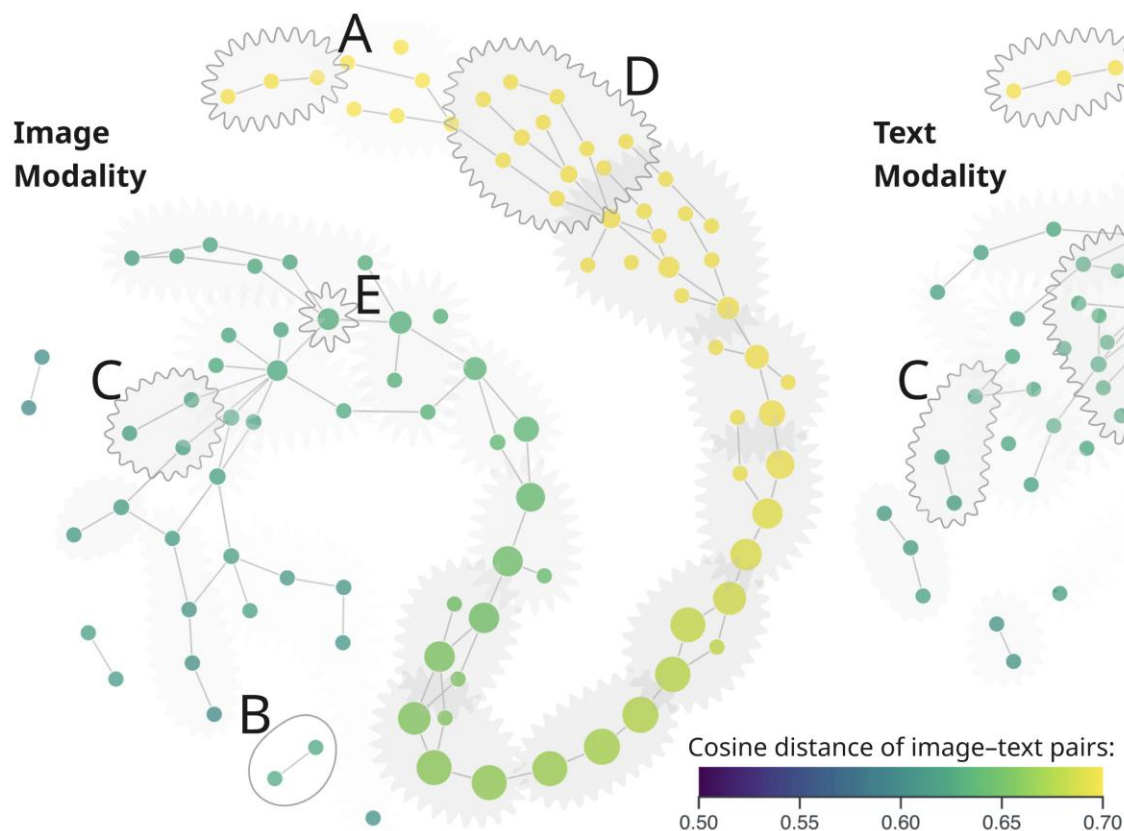
L0 - Weights / Neurons: 0 Groups, 0 Widgets

ETH zürich





Topological Alignment of Representation Spaces



APERTVS CLARITAS

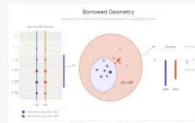
A public home for interpretability artifacts, experiments and opinions on Apertus, Switzerland's open multilingual language model.

Search

Contribute

Recent artifacts

<https://apertus-claritas.org/>



Borrowed Geometry: How Low-Resource Languages Inherit High-Resource Structure

When you ask a multilingual LLM to translate into Galician, is it actually thinking in Galician? Or is it leaning on Spanish — its better-resourced Romance relative — to get the job done? This thesis argues the latter, and provides causal evidence for it.

Technical

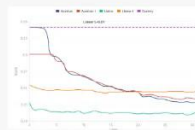
Exploratory

Mechanistic-interpretability

Multilingual

AI Translation

#apertus-1.5



Detecting and Reducing Hallucination in Multiple-Choice Questions with Activation Steering

We study mechanistic error reduction via activation steering on the Apertus family of language models, extending the Mechanistic Error Reduction with Abstention (MERA) framework to mixed-domain and cross-dataset settings.

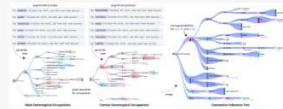
Exploratory

Steering

Mechanistic-interpretability

Active-interpretability

#apertus-1



Visualizing Hidden LLM Bias through Stochastic Path Aggregation

This post introduces TreeTracer, a visual analytics tool that aggregates hundreds of generation paths to audit Large Language Model biases. You can interact with the tool at TreeTracer.ivia.ch. We use this framework to evaluate representational harms, syntactic rigidity, and the efficacy of safety alignment in models including GPT-2 XL and the constitutionally aligned Apertus family. By developing custom visual encodings for aggregated...

Tool

Exploratory

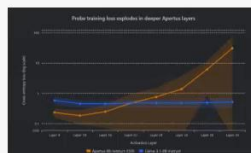
Paper

Model-evals

Visualization

Bias

#apertus-1



Active Interpretability with Hallucination Probes: Exploding Activations in Apertus-8B

Hallucination probes are small classifiers that read an LLM's internal states to flag when the model is making things up. When we trained them on 'Apertus-8B-Instruct-2509', they were far less stable than the same probes trained on 'Llama-3.1-8B-Instruct'. We traced the problem to exploding activations in deeper layers: hidden state magnitudes grow so large that the probe's optimizer can't learn a reliable decision boundary. By applying four...

Technical

Exploratory

Active-interpretability

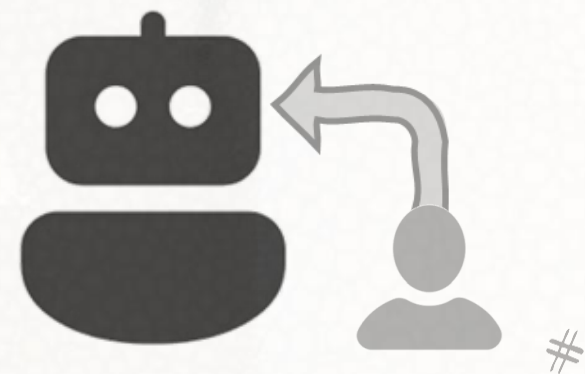
Mechanistic-interpretability

#apertus-1

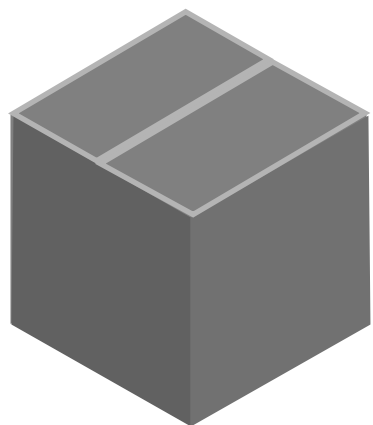


Steering Pruning the Wild Growth

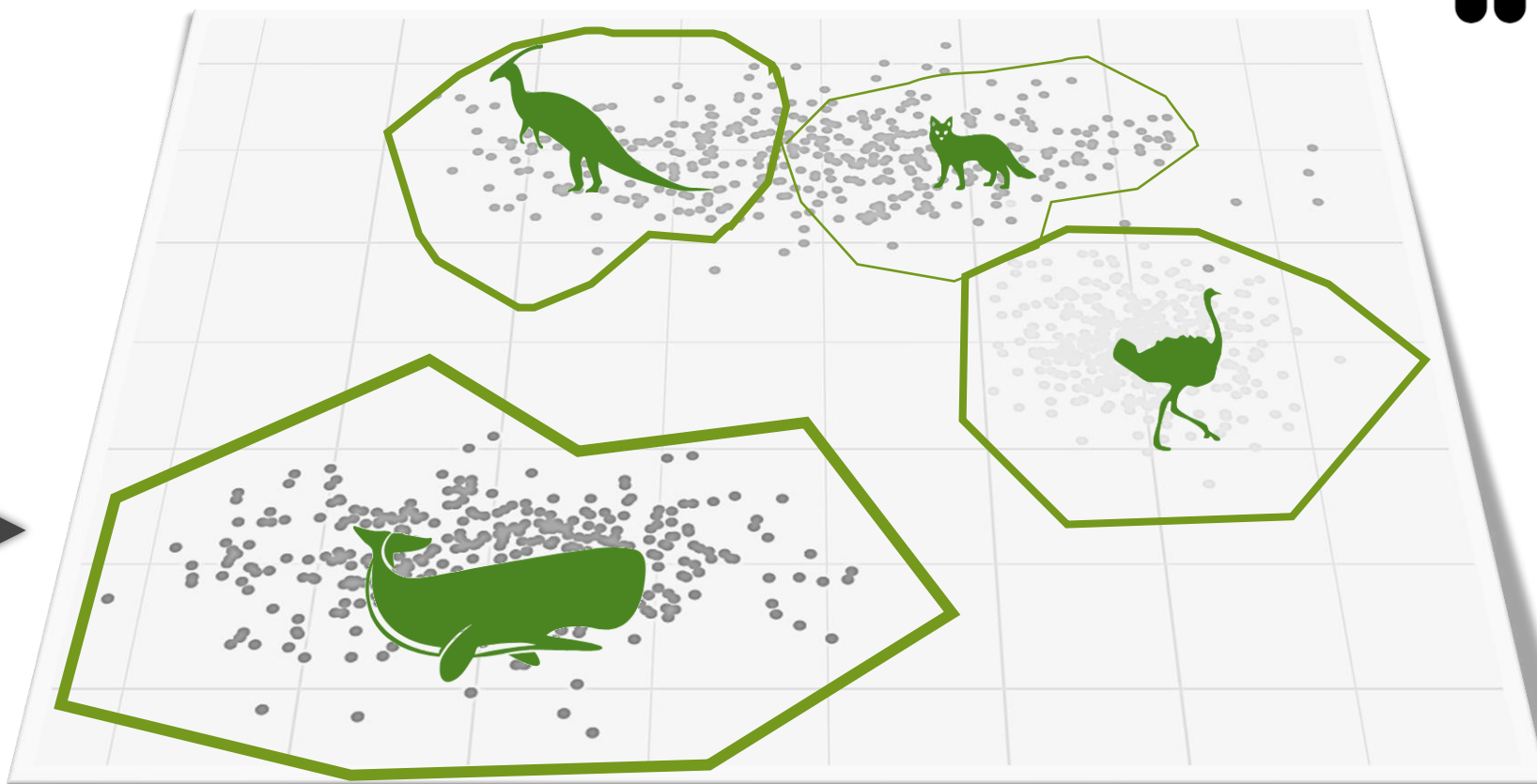
How do we guide and steer the models?



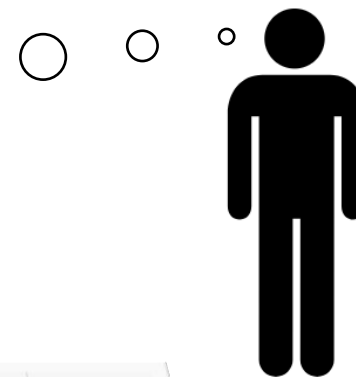
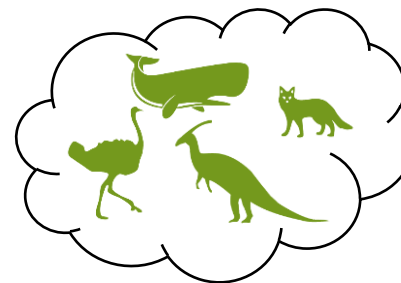
input

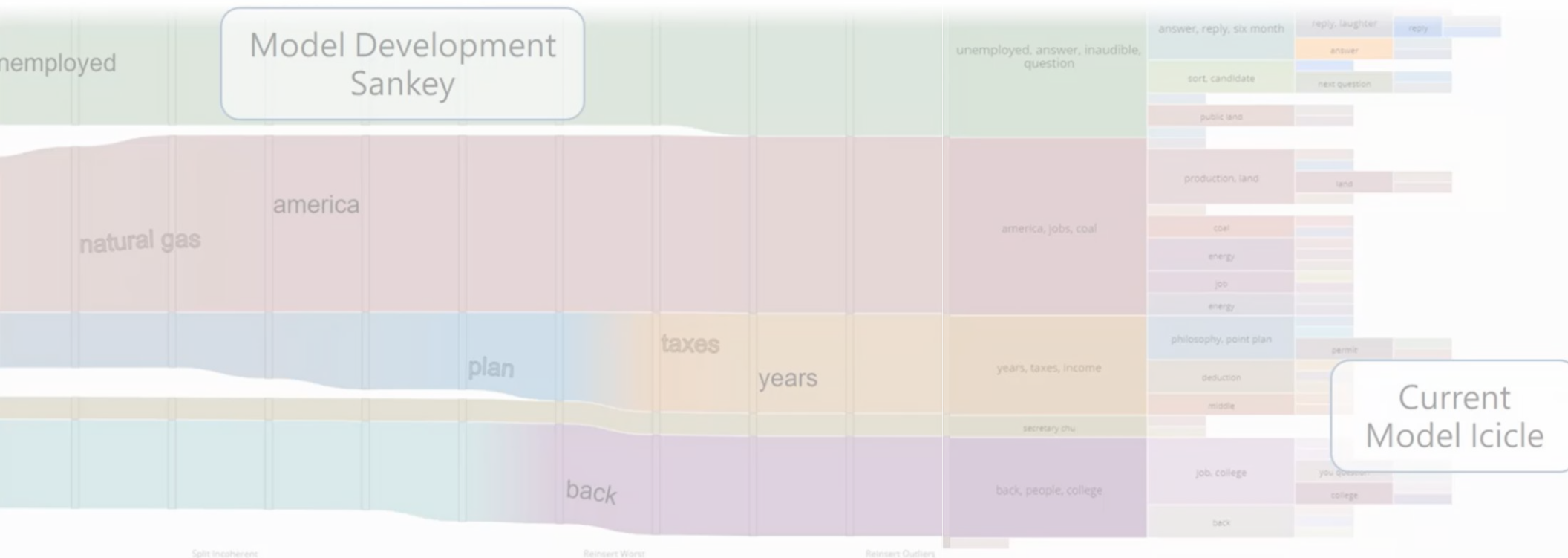


model



output





STREAM Speed MORE

Documents

Document Log

ROMNEY: You're absolutely right. And I think that's what my tax plan. The top 1 percent will continue to pay 60 percent, as they do today. I am looking to cut taxes for wealthy people. I am looking to cut taxes for middle income people. And who do I want to bring rates down, and at the same time lower exemptions and deductions, particularly for people at the high end? Because if you bring rates down, it makes it easier for small business to keep more of their capital and hire people. And for me, this is about jobs. I want to get America's economy going again. Fifty-four percent of America's workers work in businesses that are taxed as individuals. So when you bring those rates down, those small businesses are able to keep more money and hire more people. For me, I look at what's happened in the last four years and say this has been a disappointment. We can do better than this. We do n't have to settle for, how many months, 43 months with unemployment above 8 percent, 23 million Americans struggling to find a good job right now. There are 8.5 million more women living in poverty today than when the president took office. We do n't have to live like this. We can get this economy going again. My five-point plan does it. Energy independence for North America in five years. Opening up more trade, particularly in Latin America. Cracking down on China when they cheat. Getting us to a balanced budget. Fixing our training programs for our workers. And finally, championing small business. I want to make small businesses grow and thrive. I know how to make that happen. I spent my life in the private sector. I know why jobs come and why they go. And they're

Combine Similar

Alternative suggestions

Split Incoherent

Alternative suggestions

Reinsert Small

Remove Chaining

Reinsert Outliers

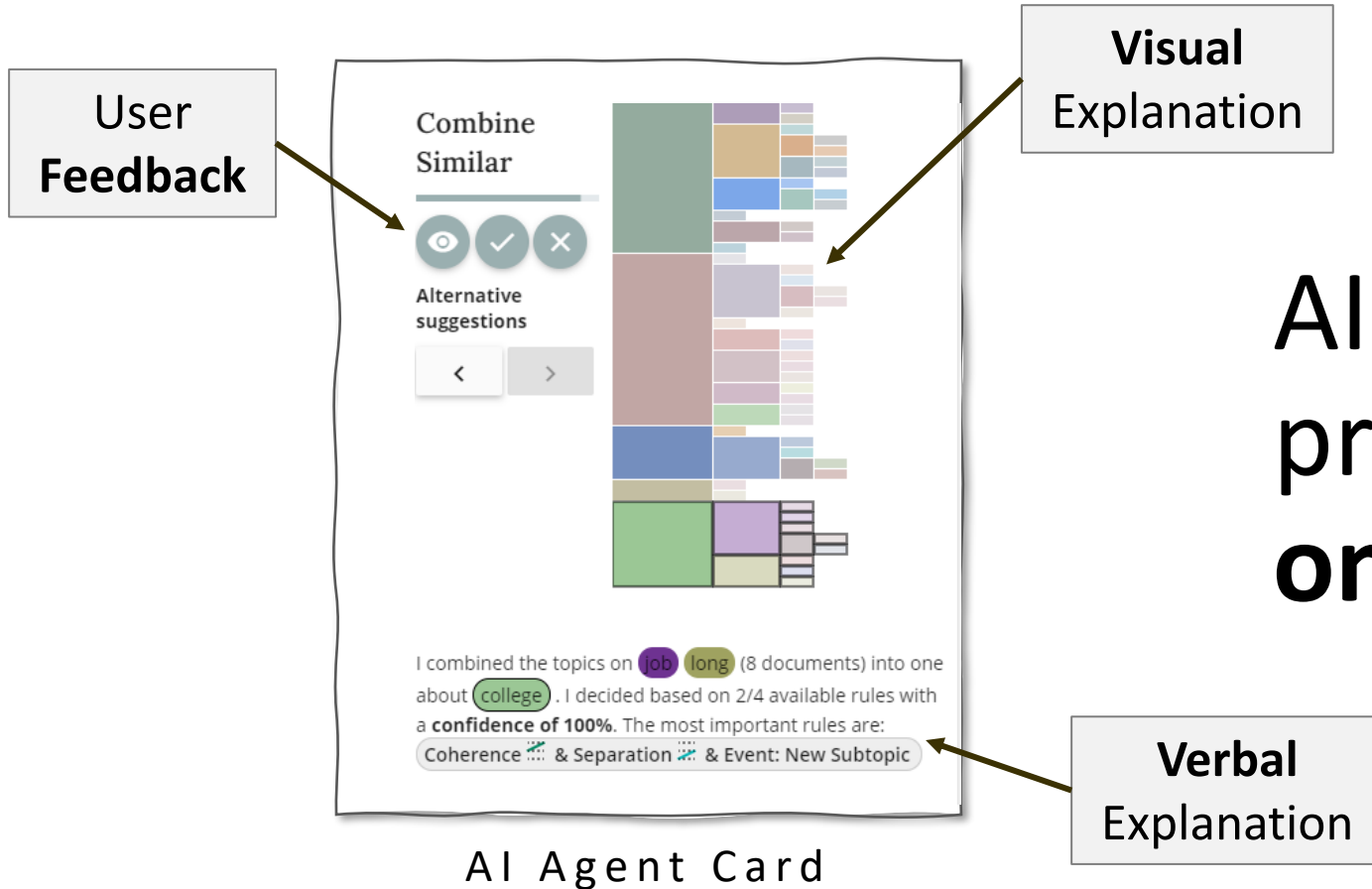
Reinsert Worst

Agent Cards

I combined the topics on back plan (13 documents) into one about college. I decided based on 2/7 available rules with a confidence of 100% and uncertainty of 71%. The most important rules are: Coherence & Separation & Event: New Subtopic

I split the topic about federal with 5 documents into two about strategy philosophy. I decided based on 4/8 available rules with a confidence of 100%. The most important rules are: Coherence & Separation & Event: New Subtopic

I removed the topic about unemployed (16 documents) as it performed worst according to my metrics. I decided based on 1/3 available rules with a confidence of 100% and uncertainty of 66%. The most important rules are: Distinctiveness & Separation & Event: New Subtopic



AI agents learn to provide suggestion  only when it is useful.



Process Control

|| STREAM Speed MORE

Documents

OBAMA:
Candy , what Governor Romney said just is n't true . He wanted to take them into bankruptcy without providing them any way to stay open . And we would have lost a million jobs . And that -- do n't take my word for it , take the executives at GM and Chrysler , some of whom are Republicans , may even support Governor Romney . But they 'll tell you his prescription was n't going to work . And Governor Romney 's says he 's got a five - point plan ? Governor Romney does n't have a five - point plan . He has a one - point plan . And that plan is to make sure that folks at the top play by a different set of rules . That 's been his philosophy in the private sector , that 's been his philosophy as governor , that 's been his philosophy as a presidential candidate . You can make a lot of money and pay lower tax rates than somebody who makes a lot less . You can ship jobs overseas and get tax breaks for it . You can invest in a company , bankrupt it , lay off the workers , strip away their pensions , and you still make money . That 's exactly the philosophy that we 've seen in place for the last decade . That 's what 's been squeezing middle class families . And we have fought back for four years to get out of that mess . The last thing we need to do is to go back to the very same policies that got us there .

ROMNEY:
Well what you 're seeing in this country is 23 million people struggling to find a job . And a lot of them , as you say , Candy , have been out of work for a long , long , long time . The president 's policies have been exercised over the last four years and they have n't put Americans back to work . We have fewer people working today than we had when the president took office . If the -- the unemployment

Combine Similar

Split Incoherent

Reinsert Small

Remove Chaining

Reinsert Outliers

Reinsert Worst

...observing & waiting to make a suggestion!

...observing & waiting to make a suggestion!

...observing & waiting to make a suggestion!

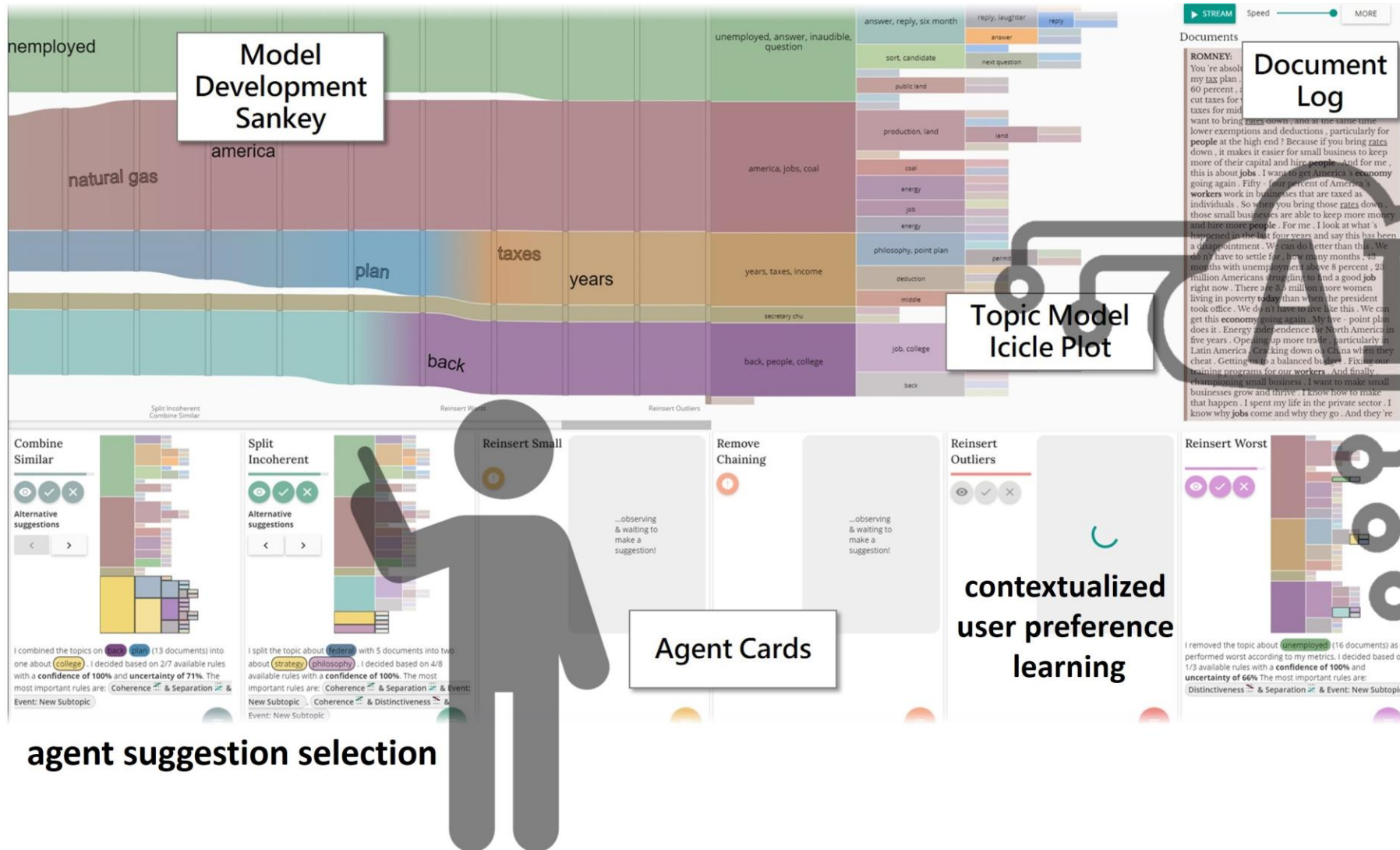
...observing & waiting to make a suggestion!

...observing & waiting to make a suggestion!

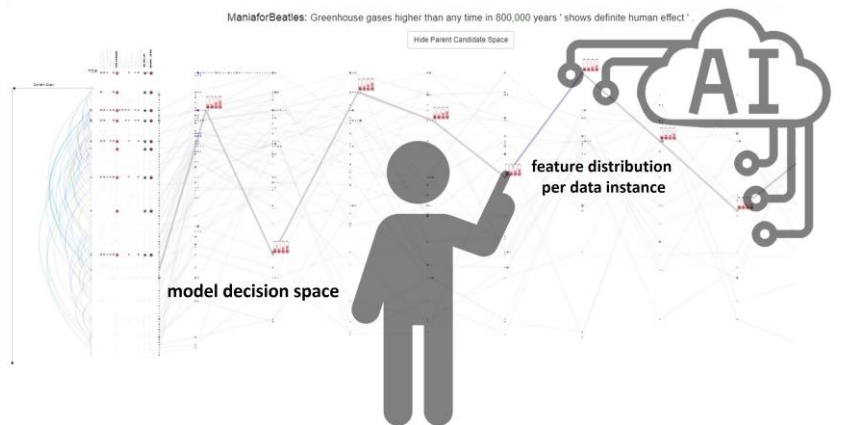
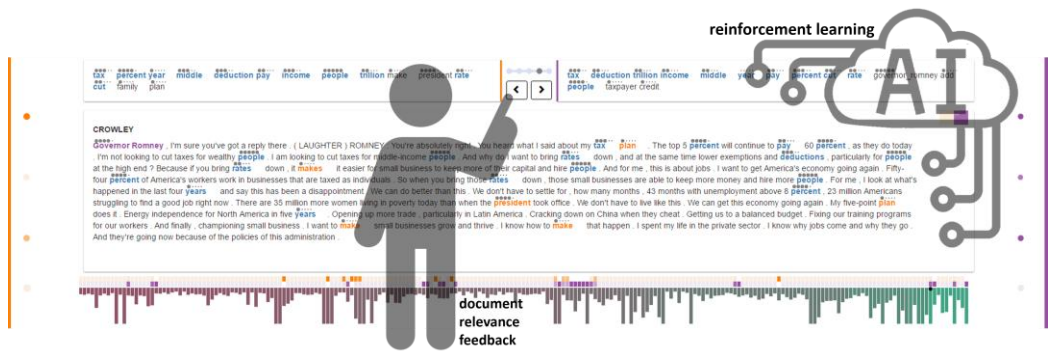
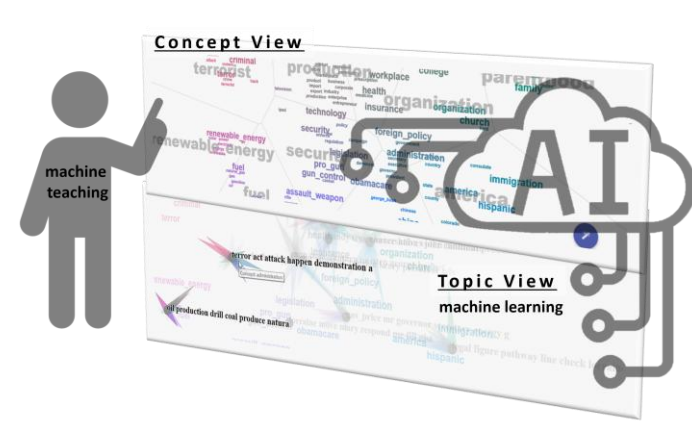
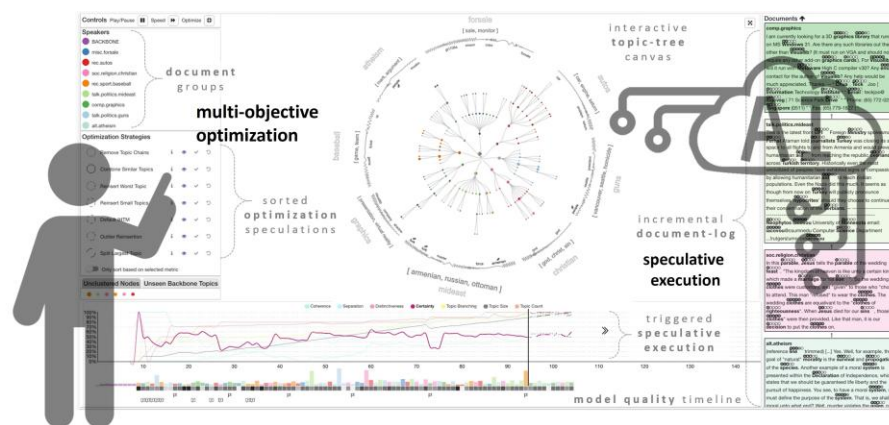
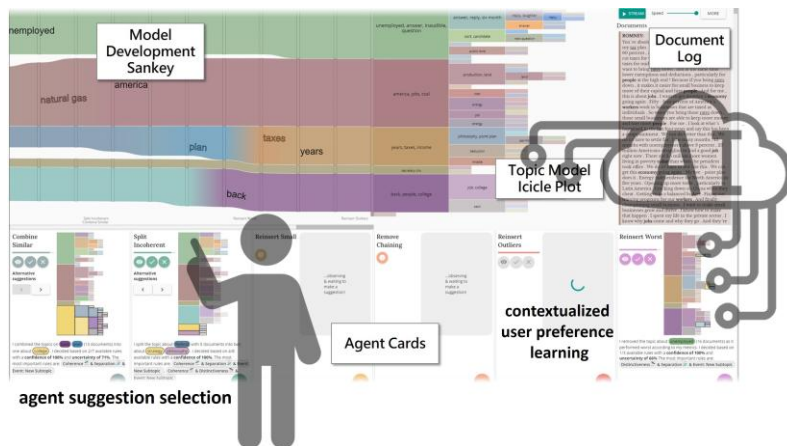
...observing & waiting to make a suggestion!

...observing & waiting to make a suggestion!

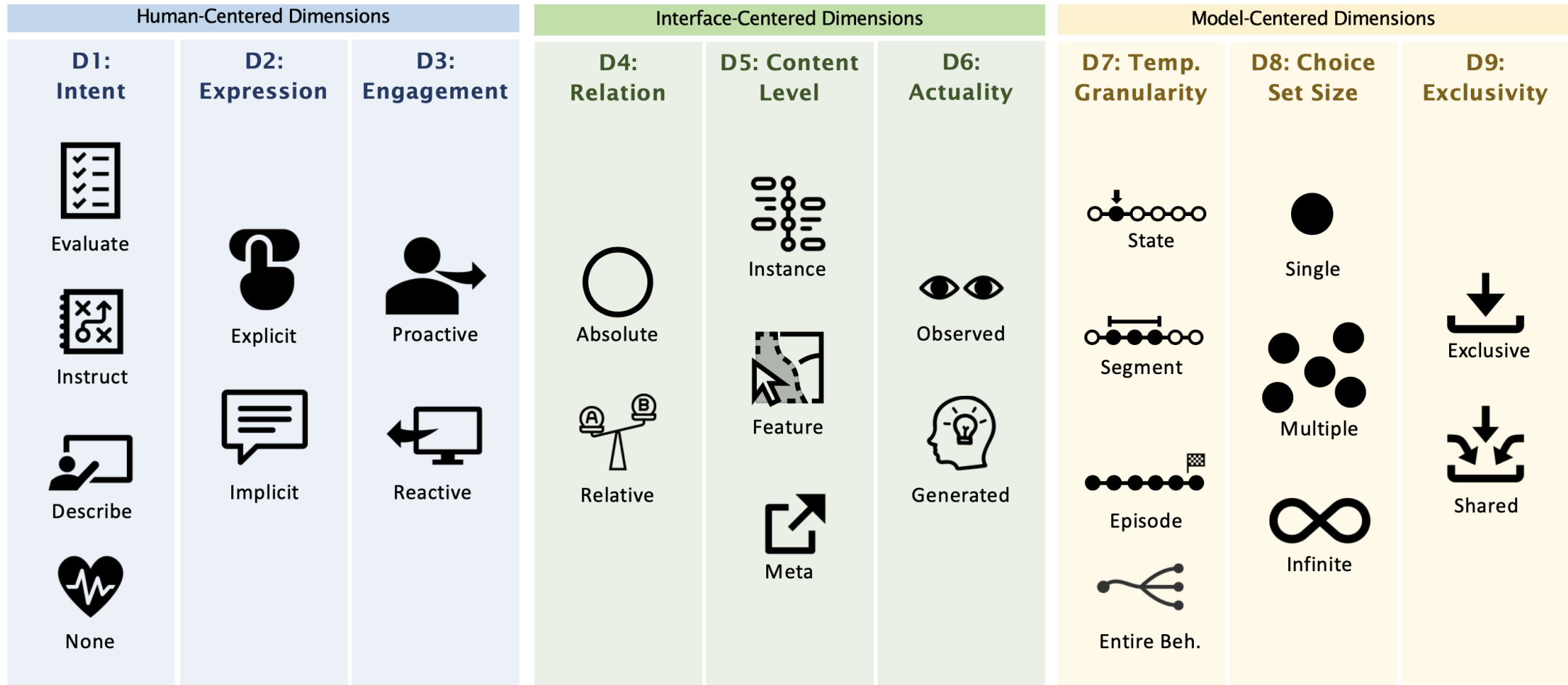
Interactive and Explainable Topic Model Refinement



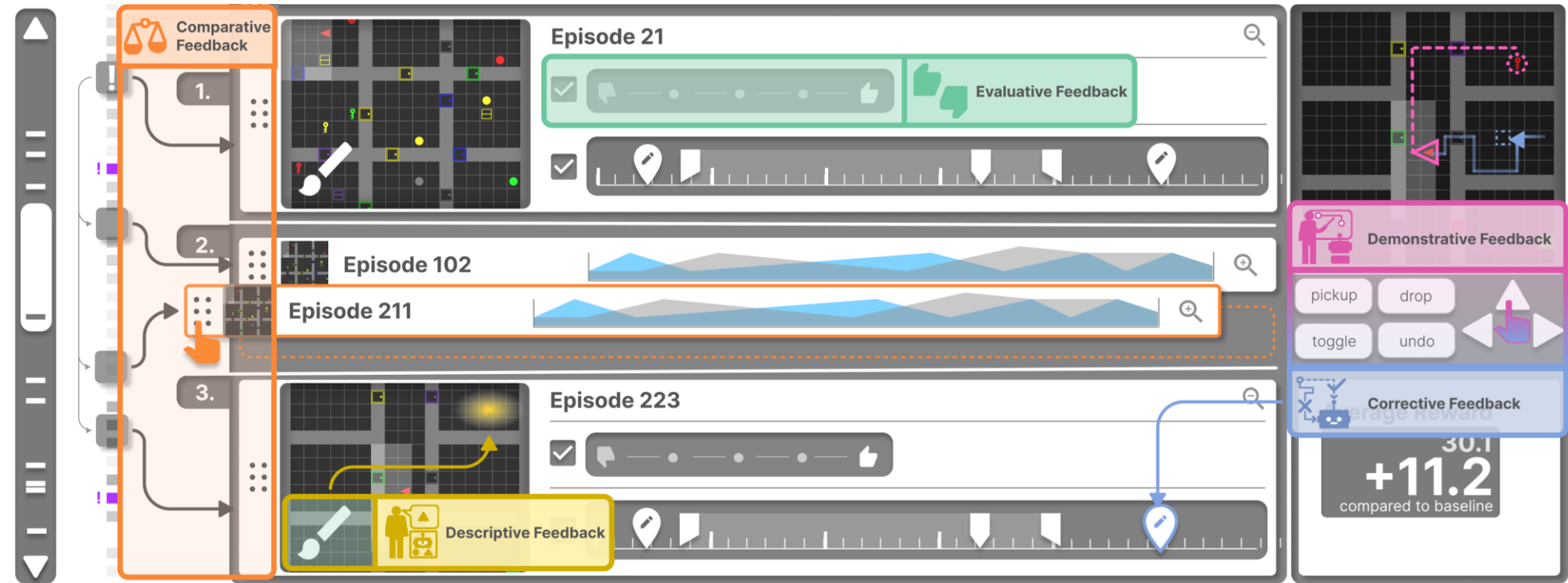
Multi-Type Feedback and Guidance



What kind of human feedback is available, i.e. how can categorize different types of human feedback?



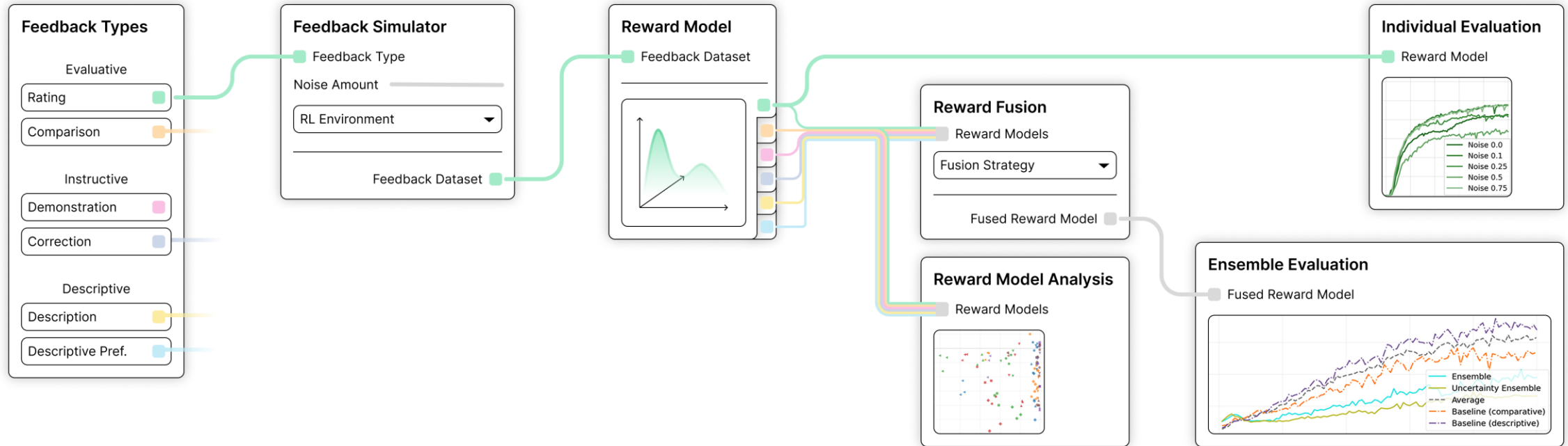
Collecting Diverse Human Feedback



RLHF-Blender: A Configurable Interactive Interface for Learning from *Diverse Human Feedback*

Y. Metz, D. Lindner, R. Baur, D. Keim, M. El-Assady, *Learning from Implicit Feedback WS at ICML2023*

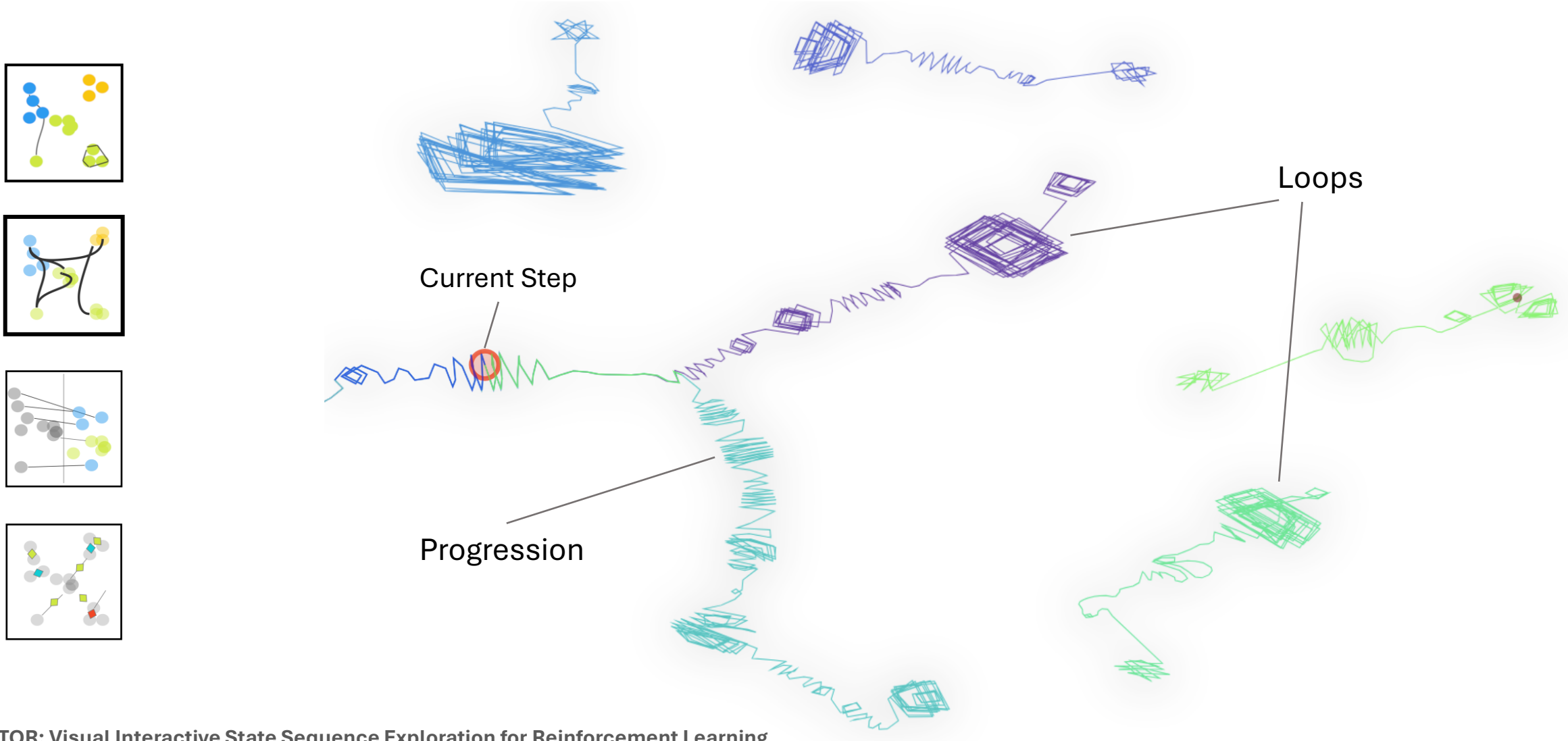
Learning from Diverse Human Feedback



Reward Learning from Multiple Feedback Types

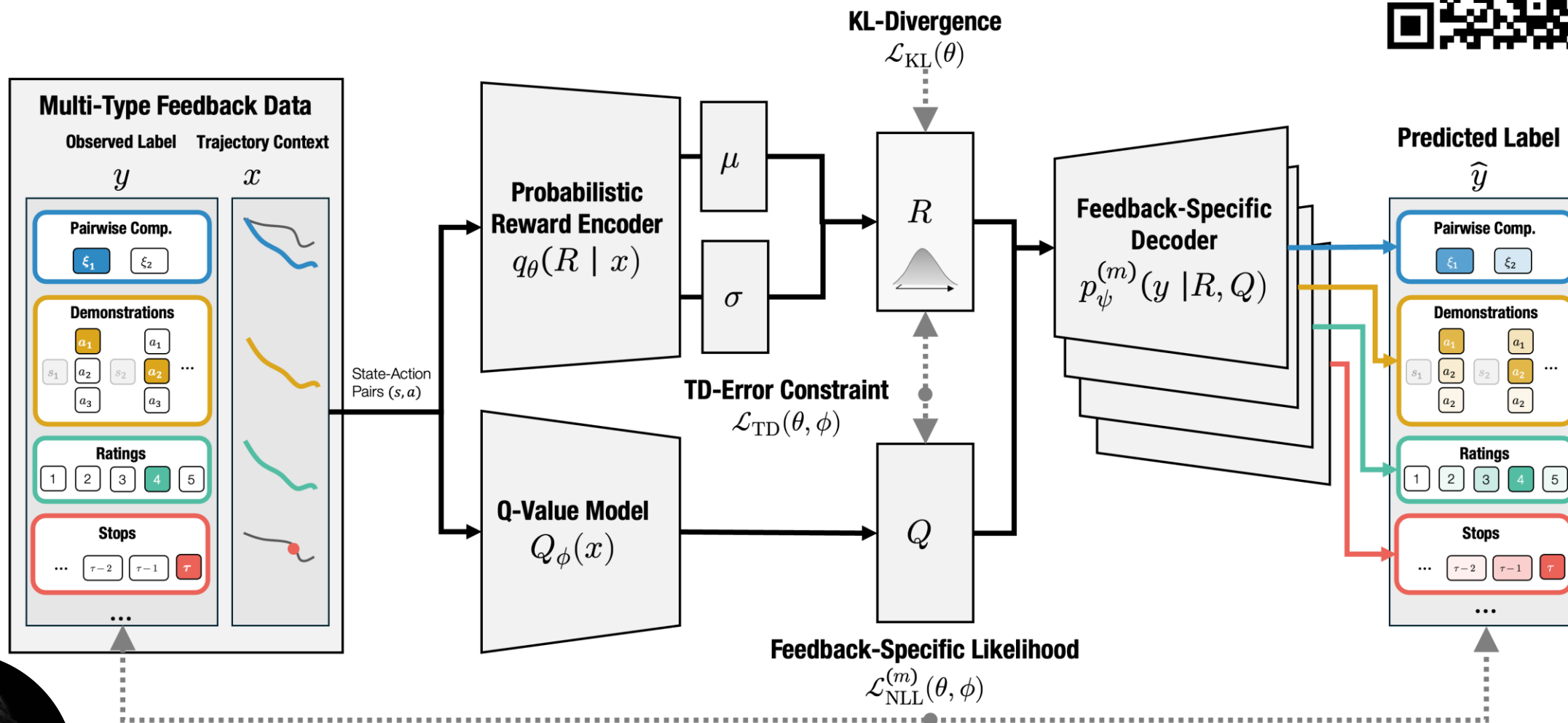
Y. Metz, A. Geiszl, R. Baur, M. El-Assady, (2025), ICLR2025

Visualizing the Agents Behavior Space





to be presented at ICML 2026



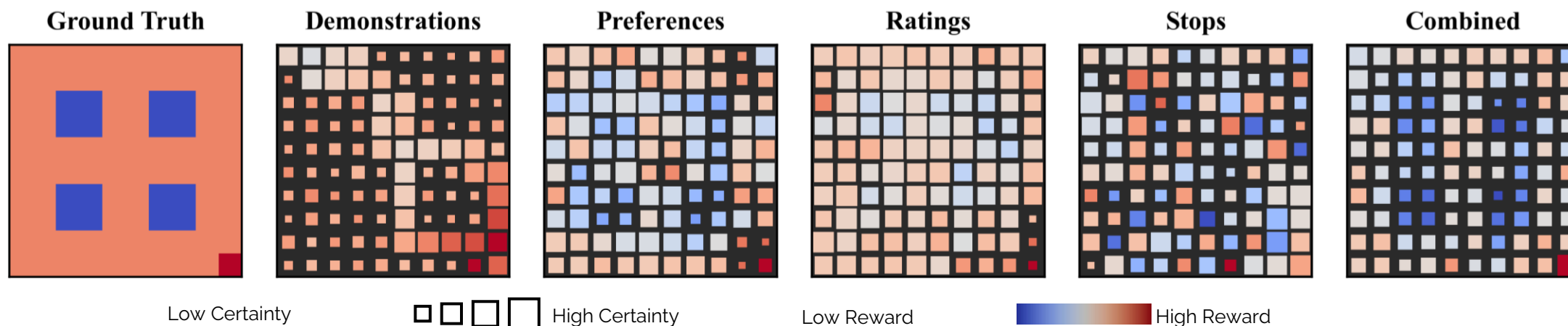
Raphaël Baur, Yannick Metz, Maria Gkoulta, Mennatallah El-Assady, Giorgia Ramponi, Thomas Kleine Buening



to be presented at ICML 2026



Learning Reward Functions from Multiple Feedback Types with Amortized Variational Inference

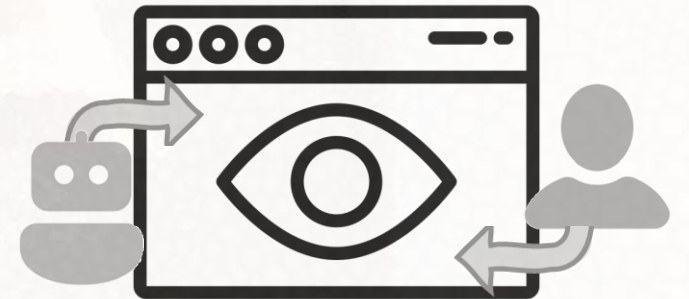


Raphaël Baur, Yannick Metz, Maria Gkoulta, Mennatallah El-Assady, Giorgia Ramponi, Thomas Kleine Buening

Accompaniment

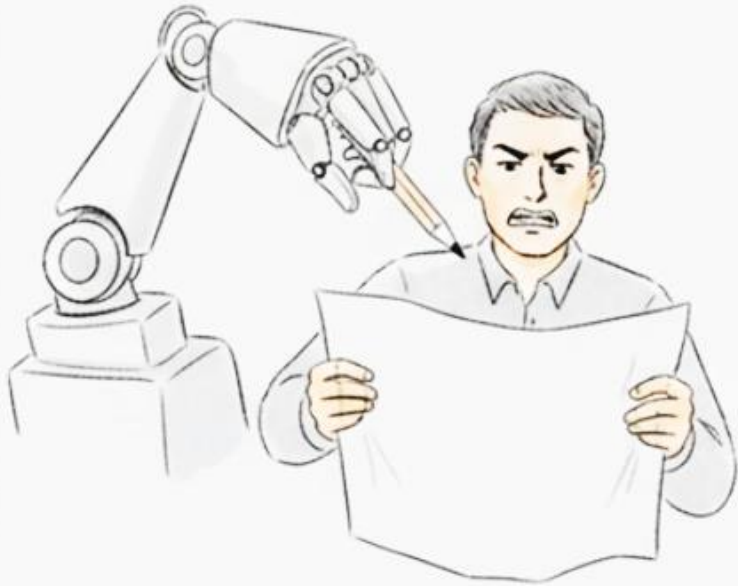
Capturing the Sweet Spot

When is the right time to interact? How do we balance agency?



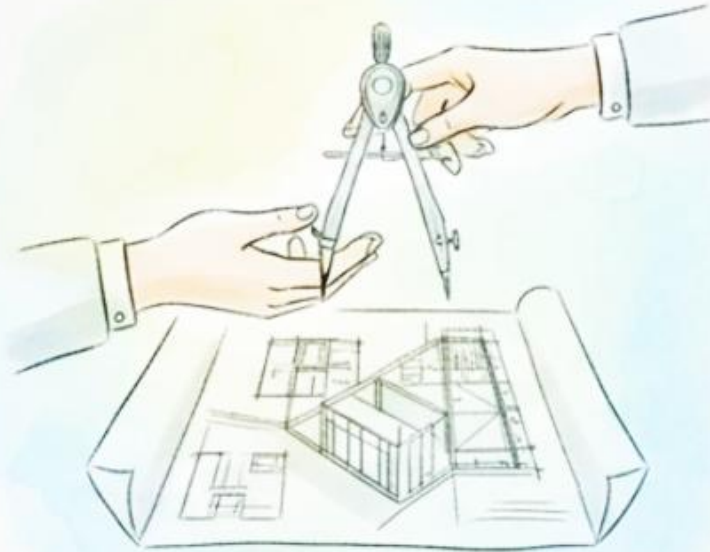
When is the right moment to interact?

Too Early



Sweet Spot

Collaborative Flow



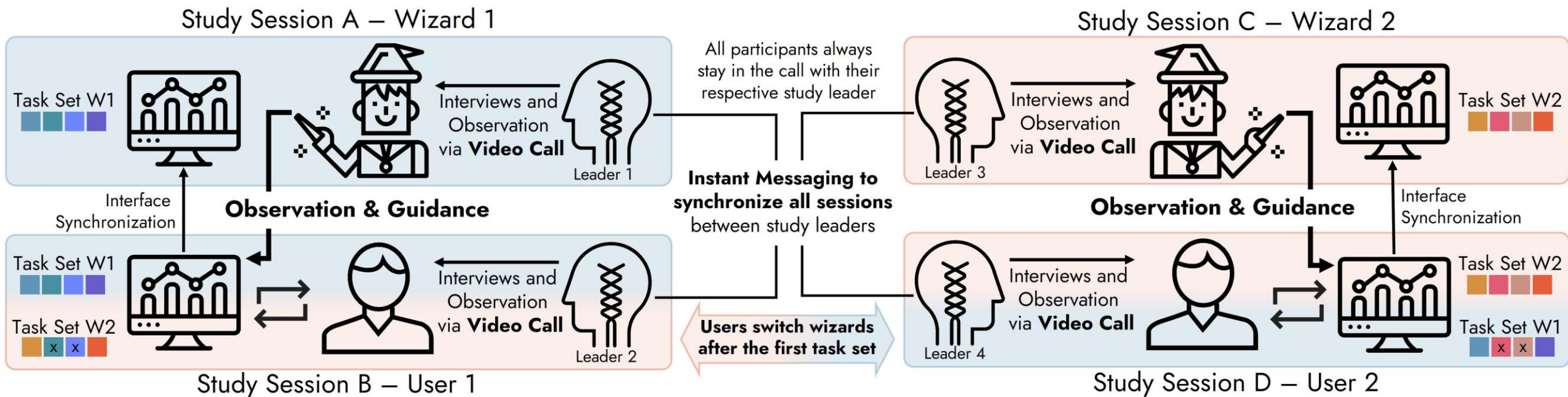
Too Late



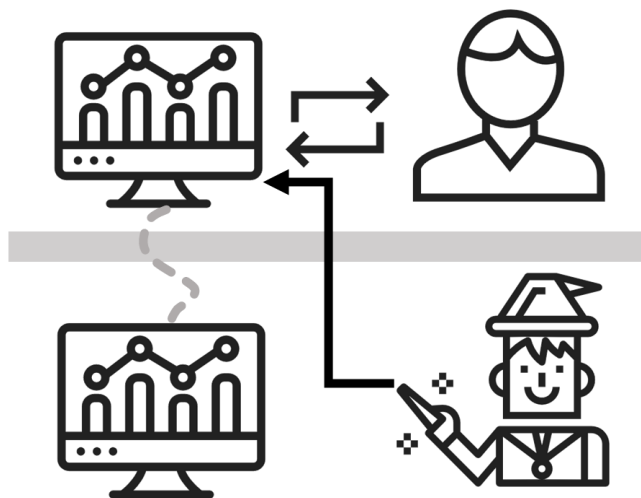
Which Guidance Strategies would Humans Perform?

Study Goals:

1. Capture guidance interaction dynamics and strategies.
2. Investigate the effects of varying guidance quality.

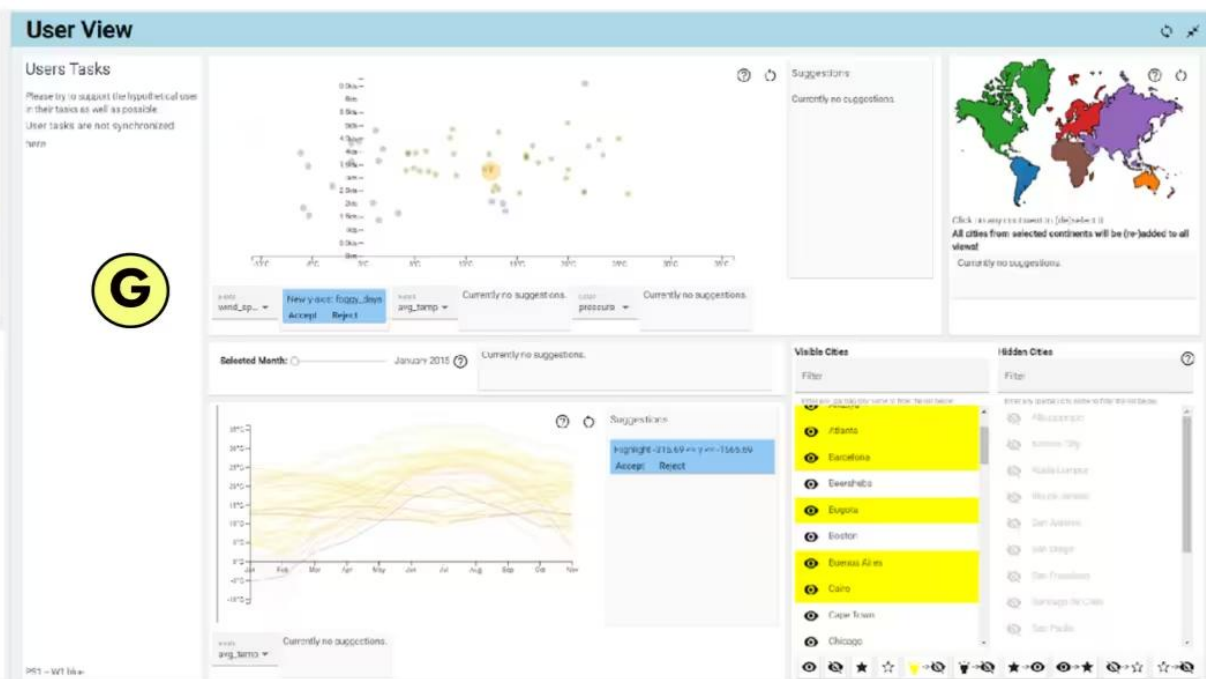
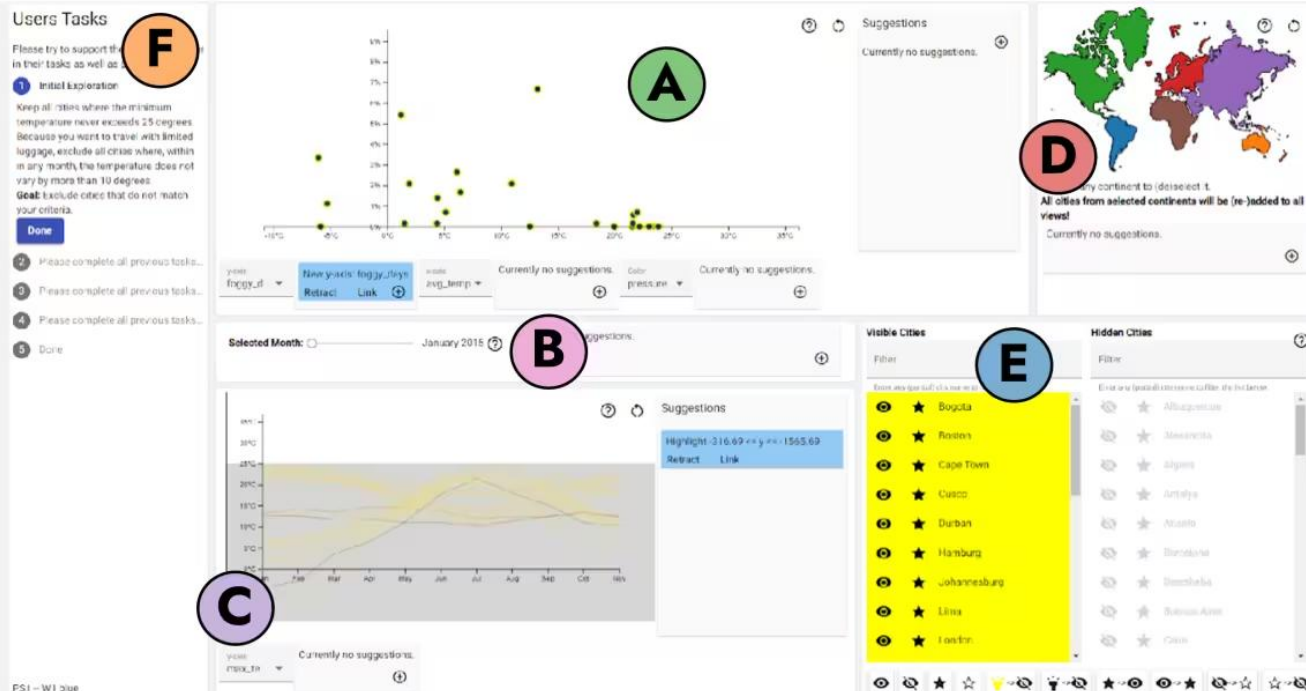
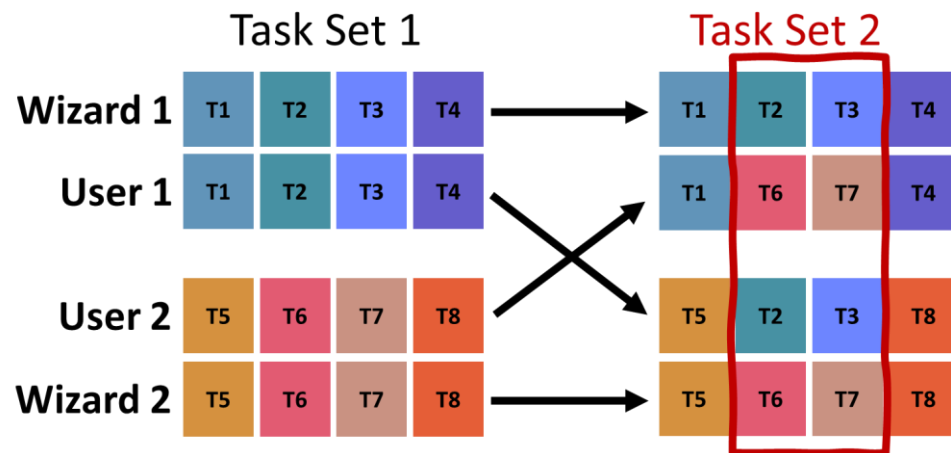


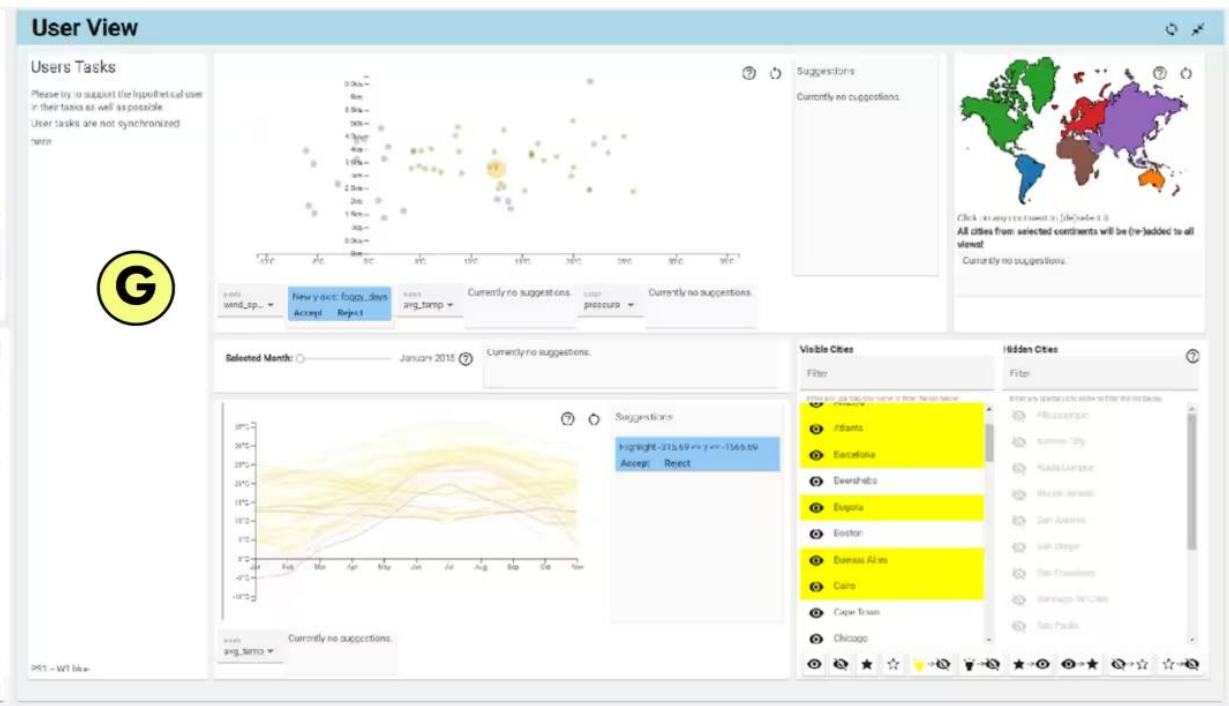
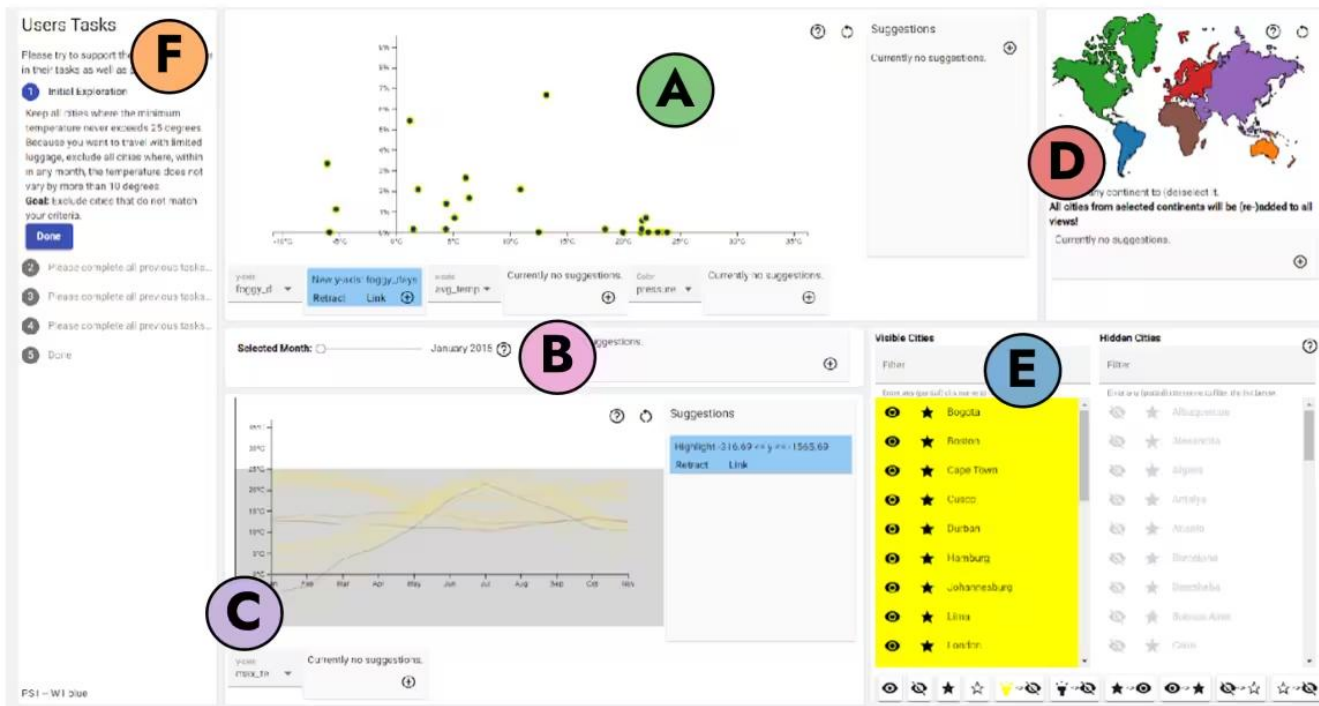
F. Sperrle, M. El-Assady, A. Arleo, D. Ceneda
A Wizard of Oz Study of Guidance Strategies and Dynamics
IEEE Transactions on Visualization and Computer Graphics, 2024

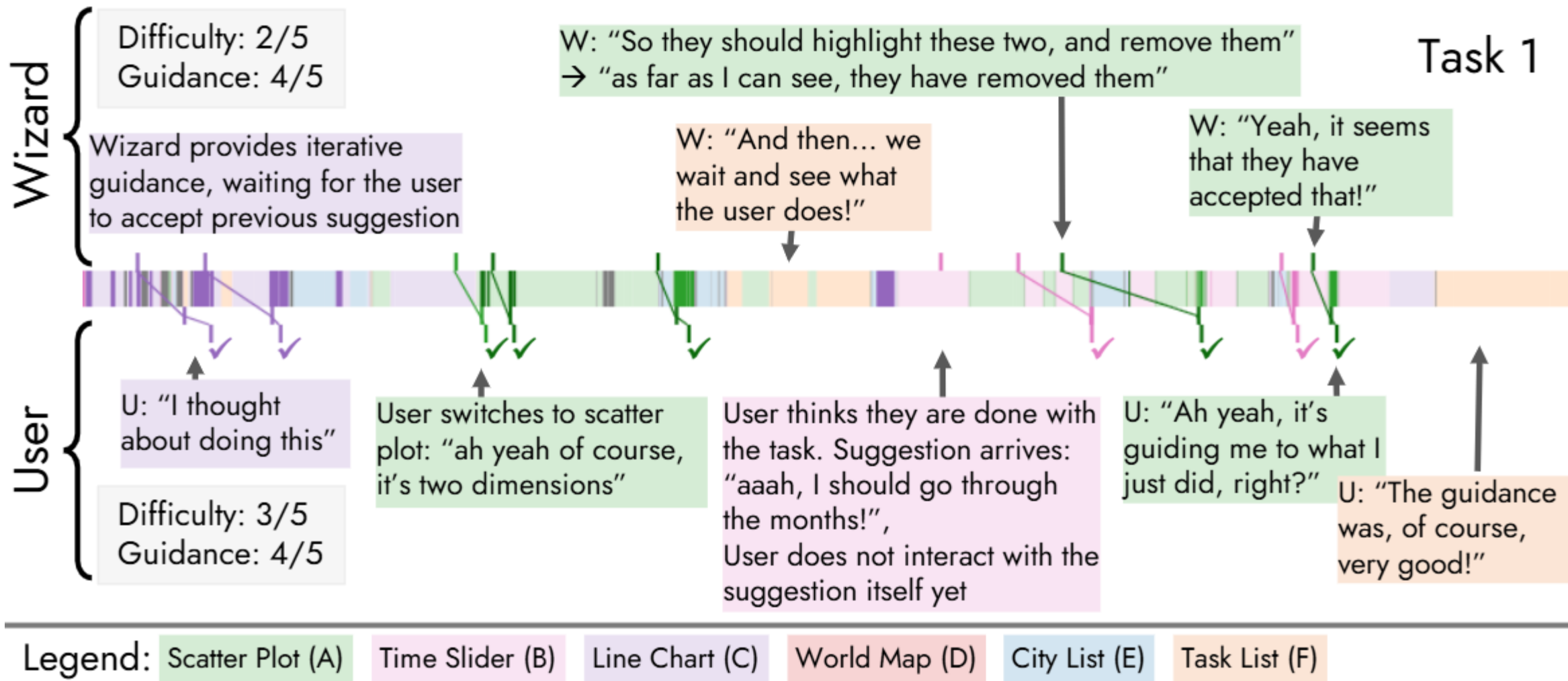


User interacts with the system to solve tasks

Wizards observes the actions and provides guidance

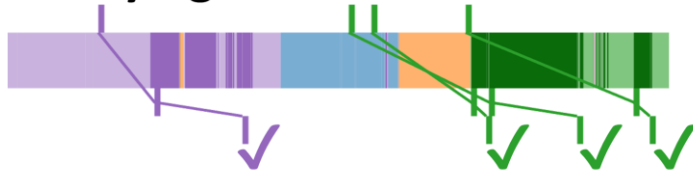






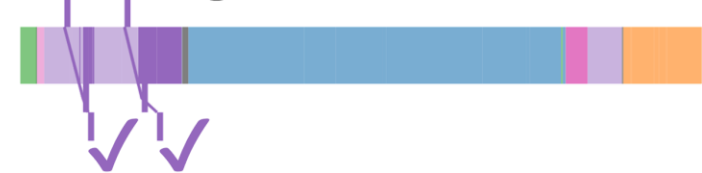
User Types

Verifying User



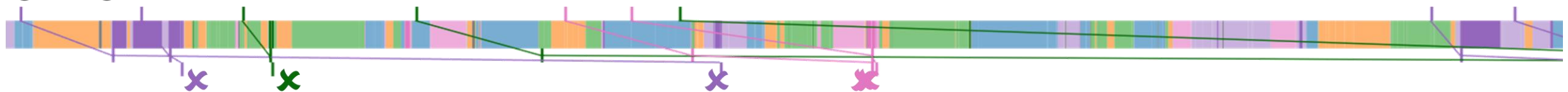
- Cautious of potential problems, spend a lot of time checking the correctness
- Benefit from suggestion justifications

Waiting for Guidance



- Intentionally pause their analysis to wait for guidance suggestions
- Depend on the suggestion

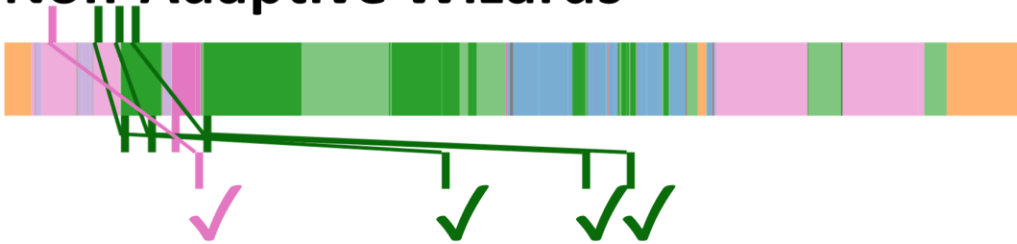
Ignoring User



- Feel that they do not need the suggestions and can complete the task independently
- Some of them were „converted“ by consistent wizards that kept making suggestions

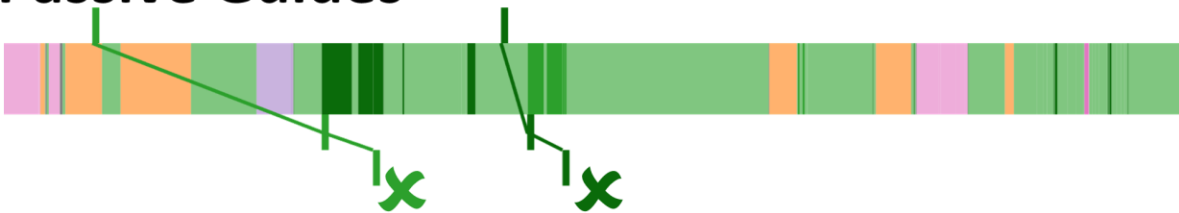
Wizard Types

Non-Adaptive Wizards



- Fast succession of suggestions, independent of user context
- Typically appreciated: no wait times for users

Passive Guides



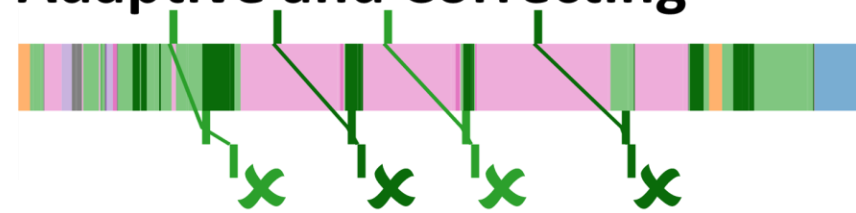
- Reduce or stop the suggestions after negative user feedback to avoid frustration
- Effectively no more guidance

Adaptive and Supporting



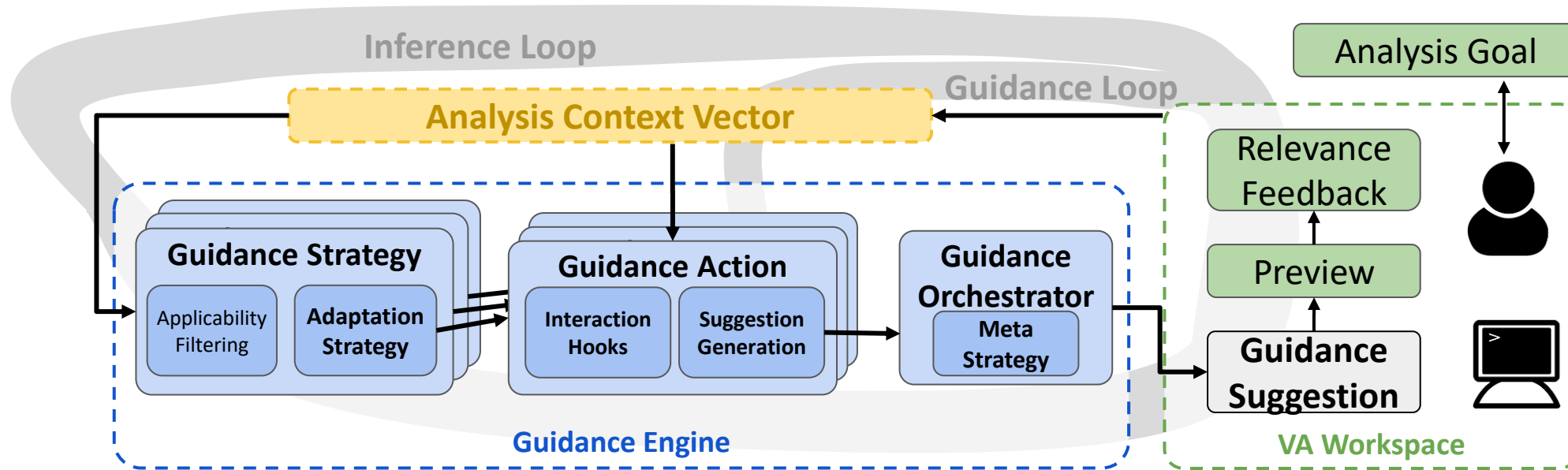
- Wizards actively monitored user progress
- Guidance suggestions only when users needed them
- Adjust suggestion content to user reaction

Adaptive and Correcting



- Wizards actively monitored user progress
- Wizards try to convince users of a different analysis path

Lotse: Strategy-Based Guidance Model



Implementation-centred model of guidance

 <https://github.com/lotse-guidance>  <https://pypi.org/project/Lotse/>

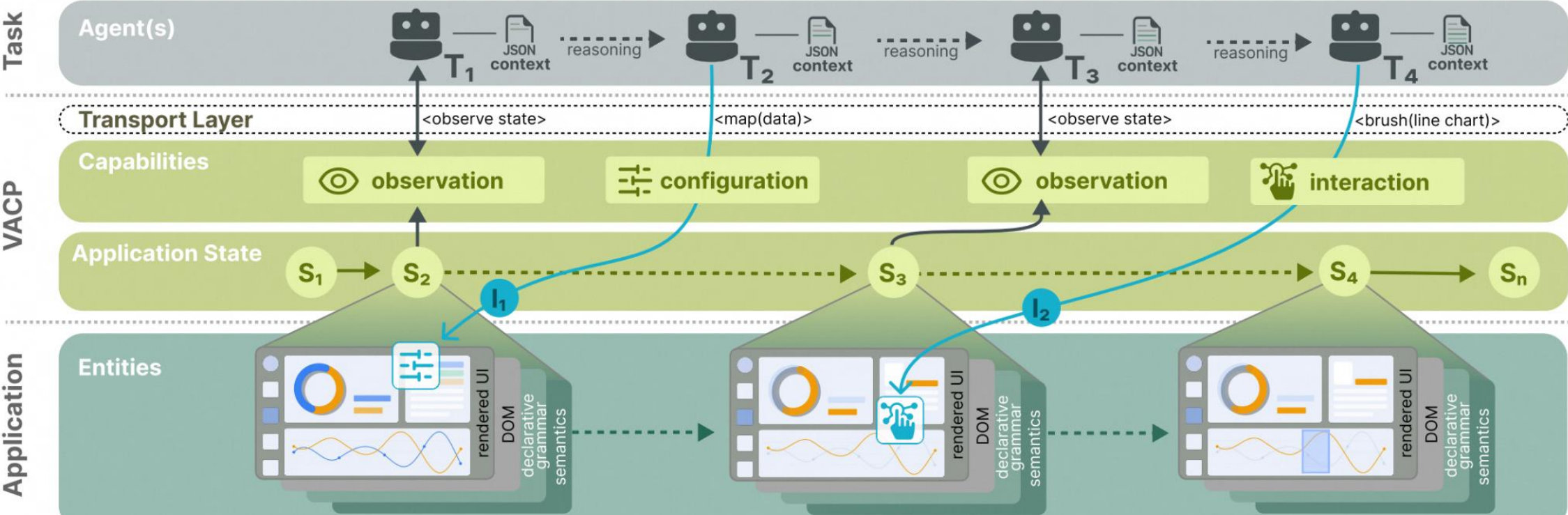
Integration examples for d3, vega, angular, react

Visual Analytics Context Protocol

A small in-app contract for state snapshots and semantic actions in visual analytics.

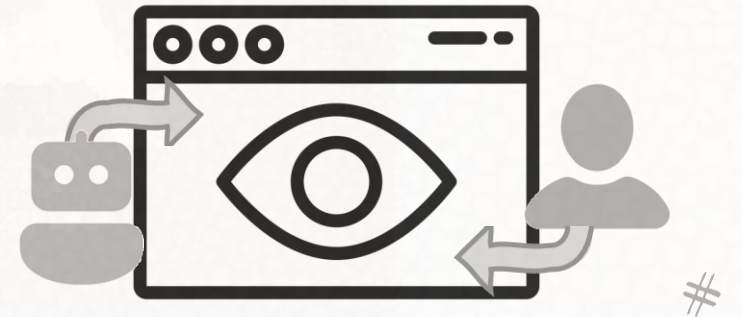


<https://vacp.dev/>



Grounding the Process

How do we externalize *tacit knowledge*?



Do We Have the Knowledge We Need?

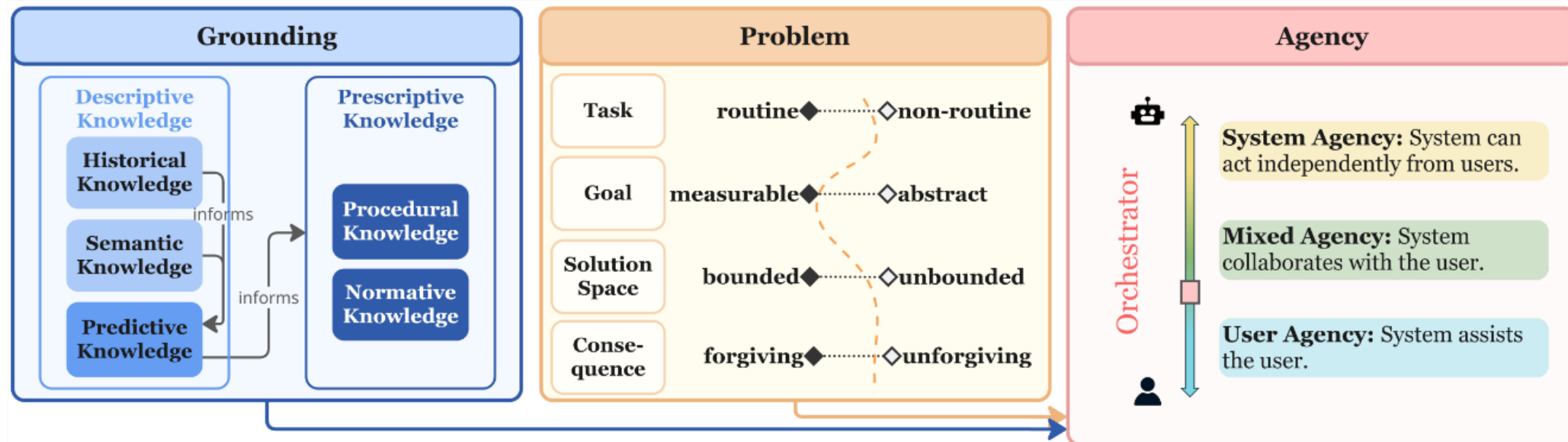
Rethinking Human-AI Decision-Making in Corporations

ANNE MARX, Department of Computer Science & ETH AI Center, ETH Zurich, Switzerland

RICARDO MAIA AVELINO, Department of Computer Science & Architecture, ETH Zurich, Switzerland

TORBJØRN NETLAND, Department of Management, Technology, and Economics, ETH Zurich, Switzerland

MENNATALLAH EL-ASSADY, Department of Computer Science, ETH Zurich, Switzerland





read the paper

Pragmatic Repair in Conversations



Principle of Least Collaborative Effort

Humans aim to minimize the conversational effort by

- Initially using **vague** statements.
- **Iteratively** asking for clarification to disambiguate **meaning, prioritizing** most uncertain points.



read the paper

Automated Pragmatic Repair

What was the review of the drama film?

1. *Generate
Interpretations*



```
SELECT Opinion  
FROM Film  
WHERE Genre = 'Drama'
```

```
SELECT Opinion  
FROM Film  
WHERE Genre LIKE '%Drama'
```

```
SELECT F.Opinion  
FROM Film F  
WHERE F.Genre = 'Drama'
```

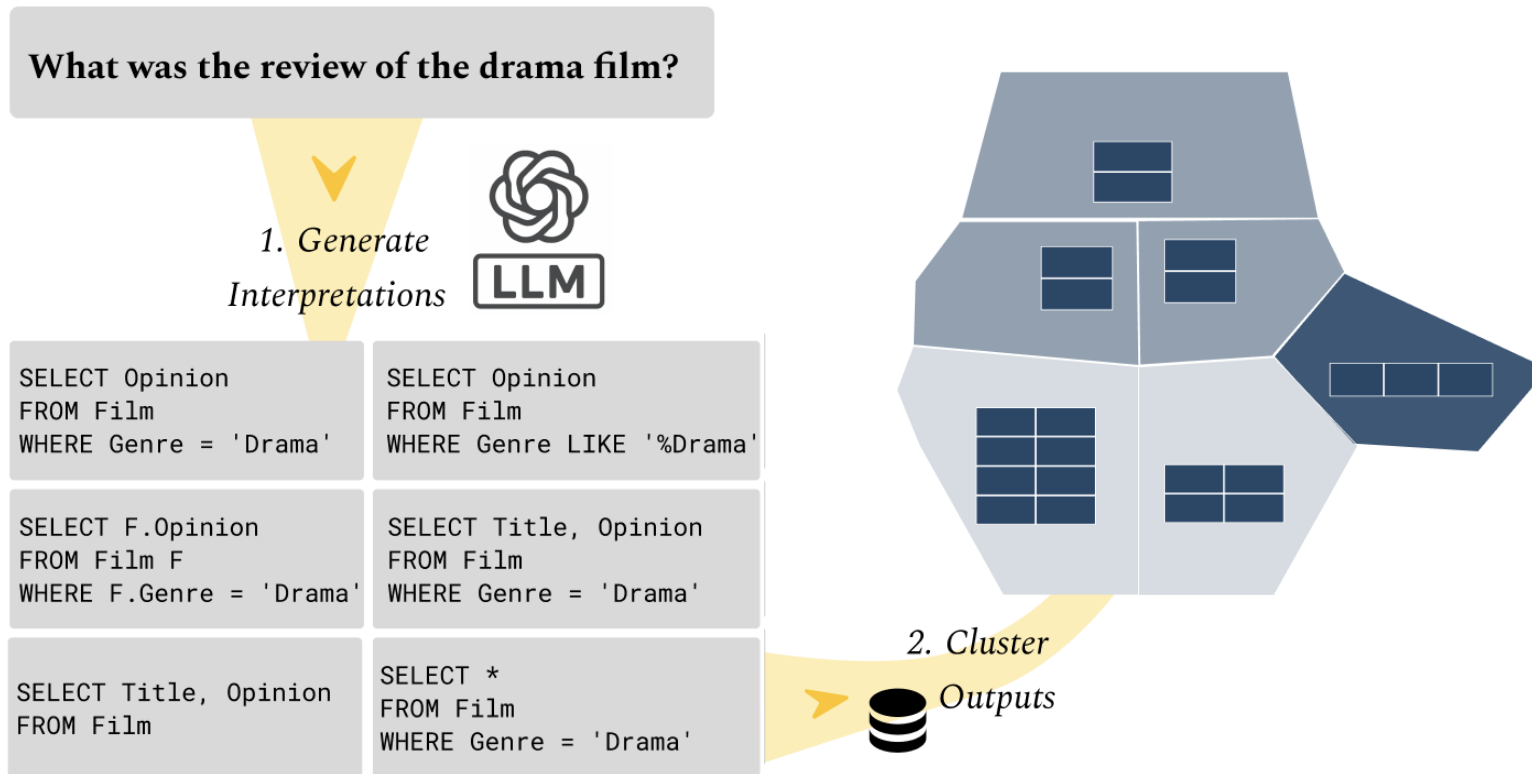
```
SELECT Title, Opinion  
FROM Film  
WHERE Genre = 'Drama'
```

```
SELECT Title, Opinion  
FROM Film
```

```
SELECT *  
FROM Film  
WHERE Genre = 'Drama'
```

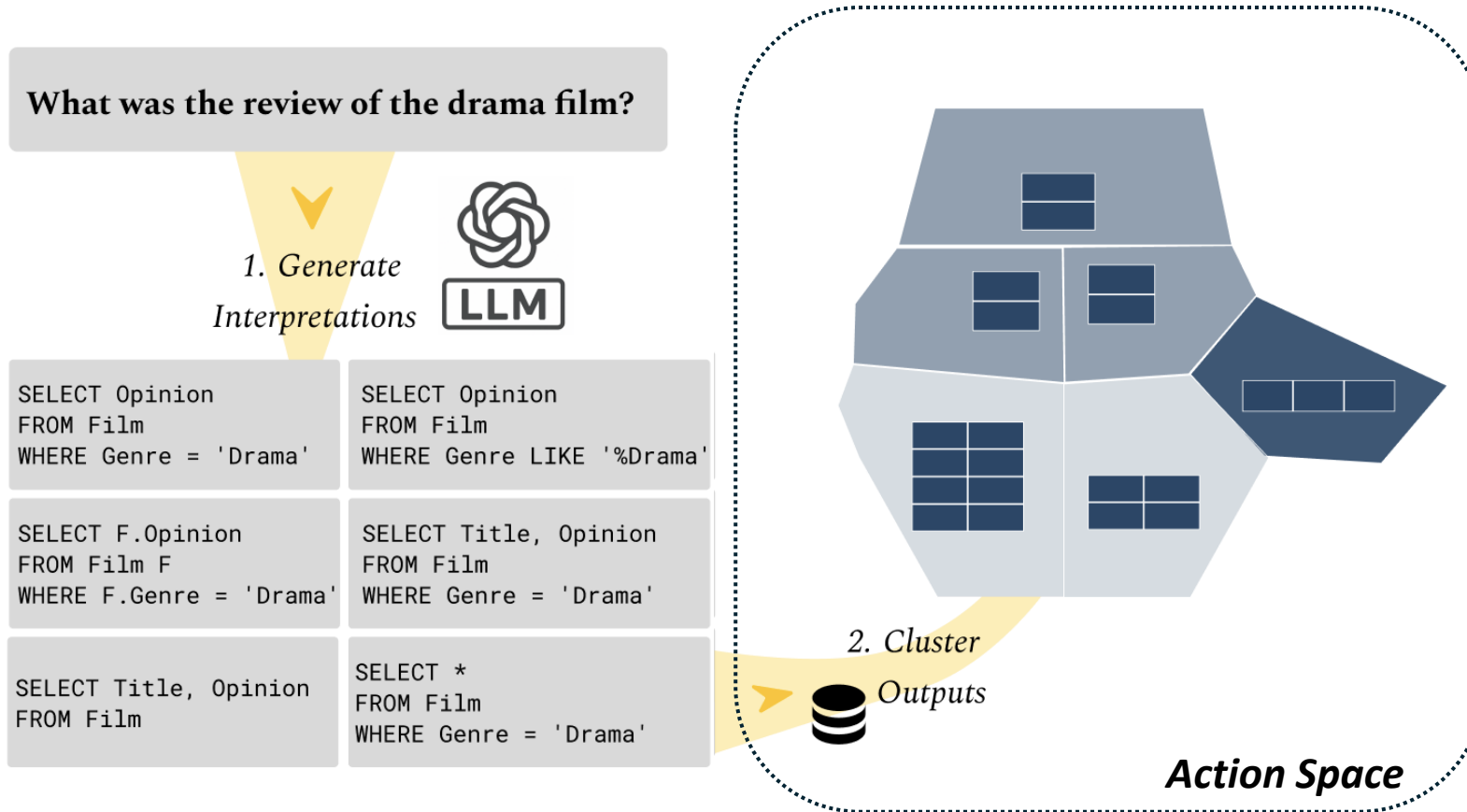



Automated Pragmatic Repair



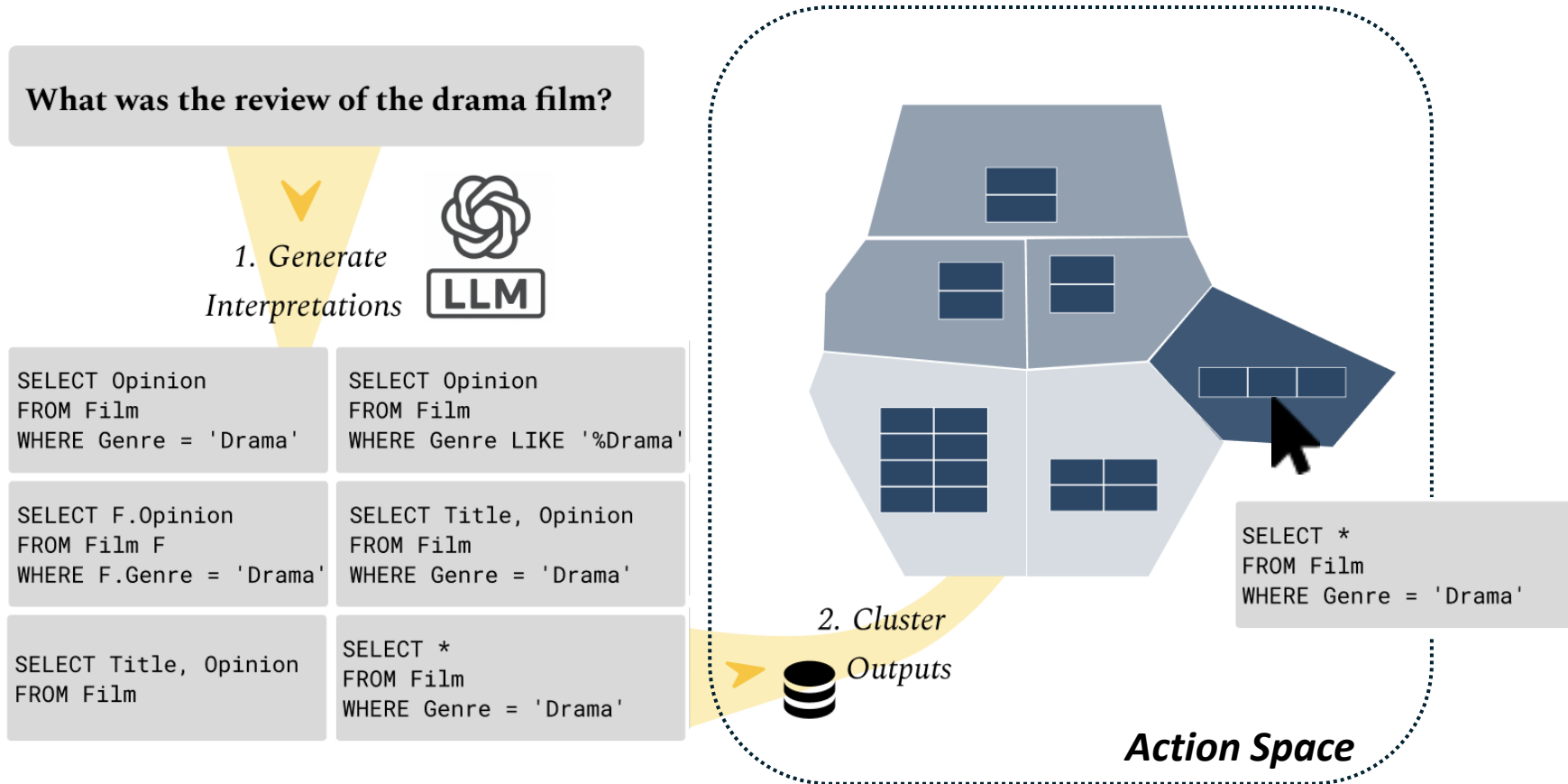


Automated Pragmatic Repair



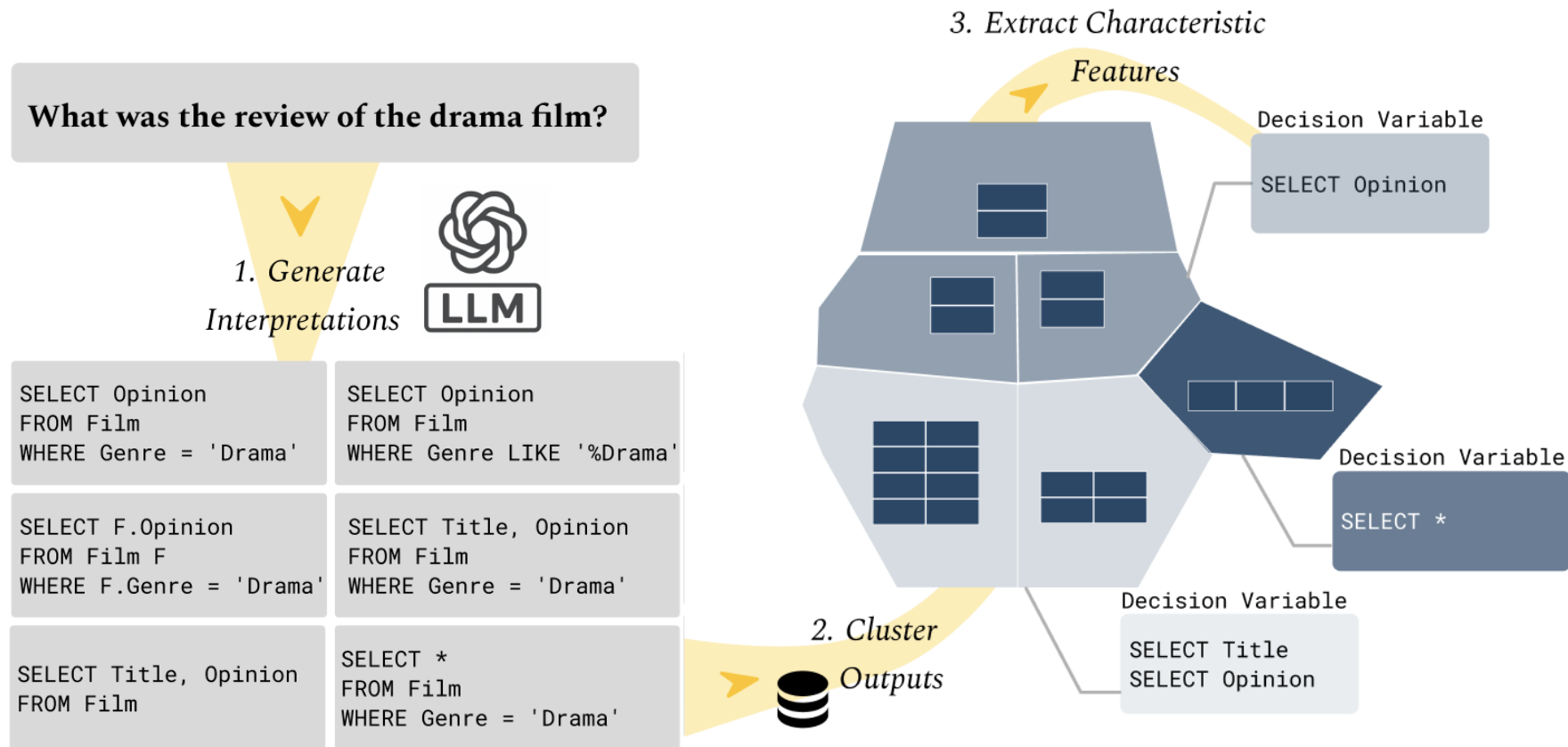


Automated Pragmatic Repair



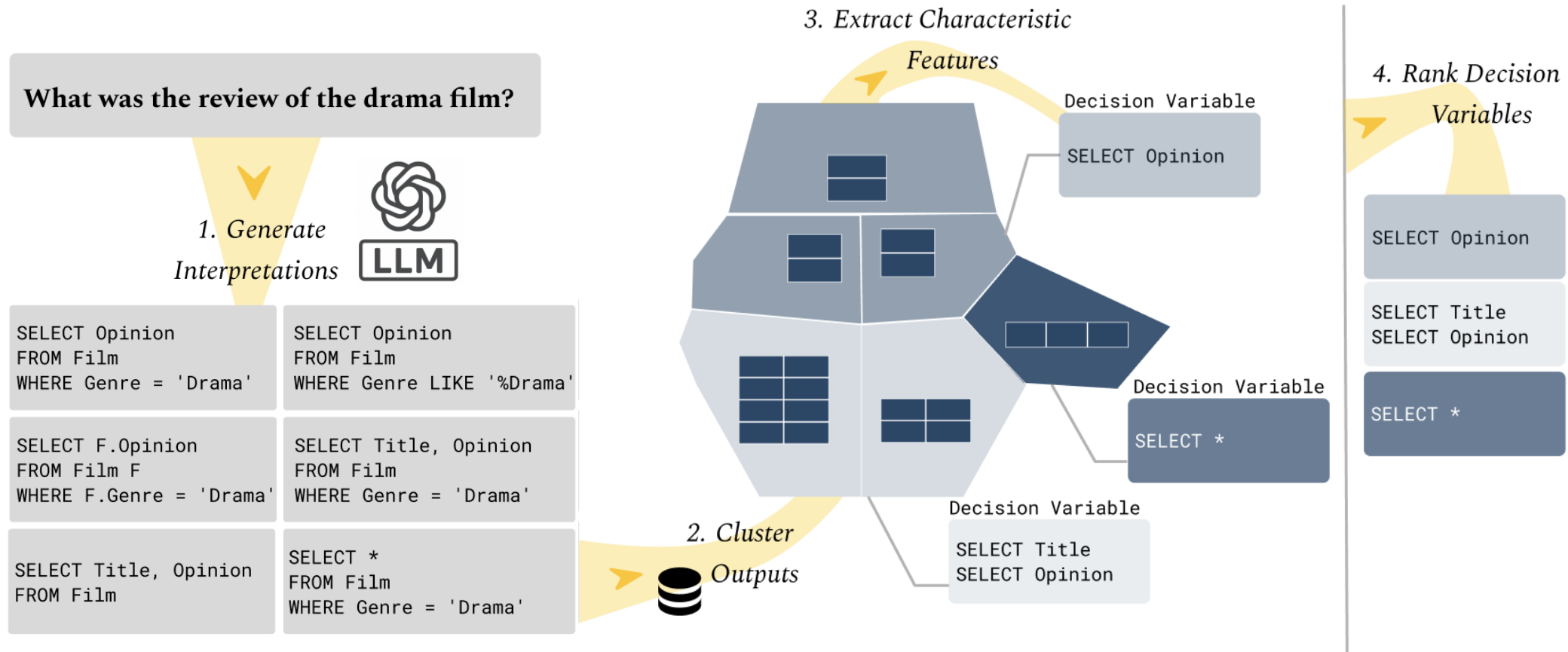


Automated Pragmatic Repair





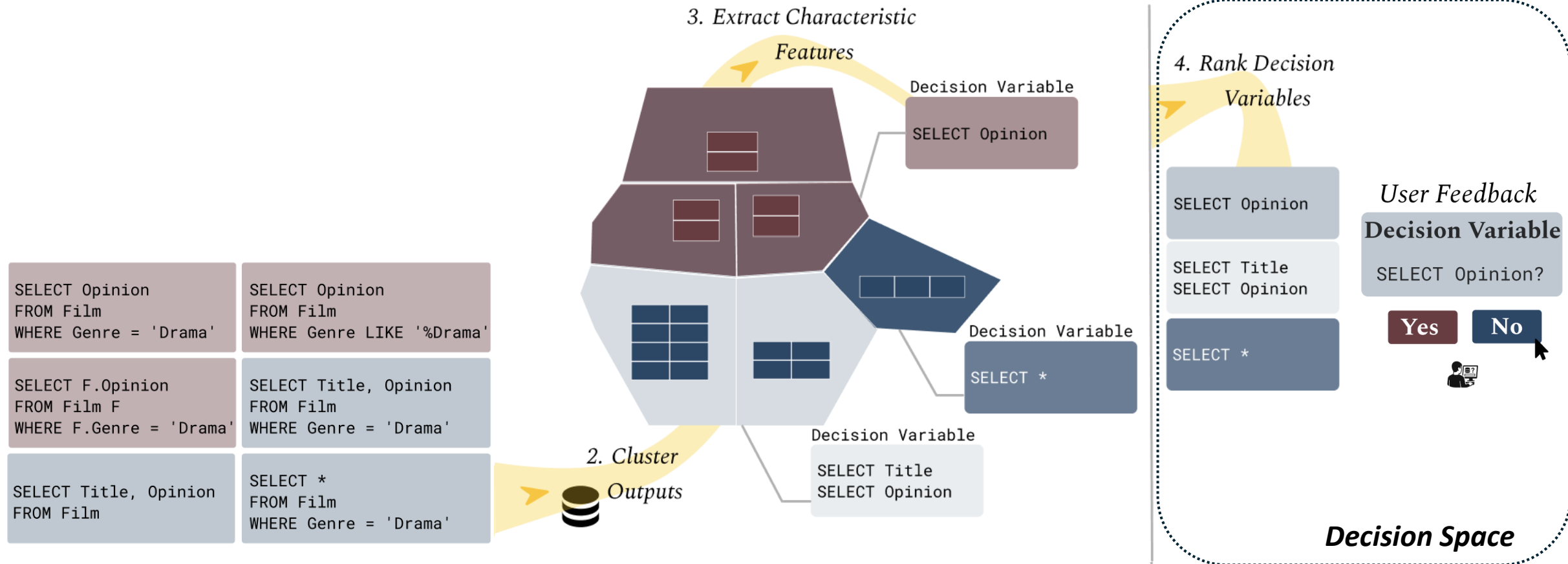
Automated Pragmatic Repair



Automated Pragmatic Repair



read the paper

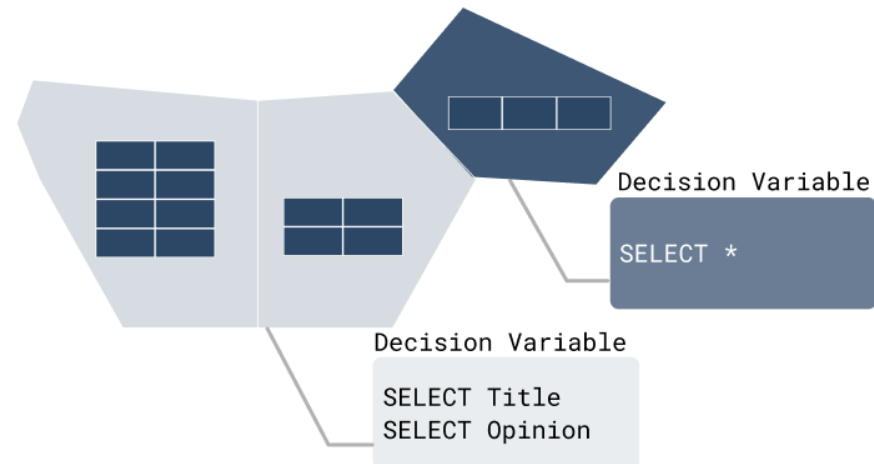




read the paper

Automated Pragmatic Repair

SELECT Opinion FROM Film WHERE Genre = 'Drama'	SELECT Opinion FROM Film WHERE Genre LIKE '%Drama'
SELECT F.Opinion FROM Film F WHERE F.Genre = 'Drama'	SELECT Title, Opinion FROM Film WHERE Genre = 'Drama'
SELECT Title, Opinion FROM Film	SELECT * FROM Film WHERE Genre = 'Drama'



4. Rank Decision
Variables

SELECT Opinion

SELECT Title
SELECT Opinion

SELECT *

User Feedback
Decision Variable
SELECT Opinion?

Yes

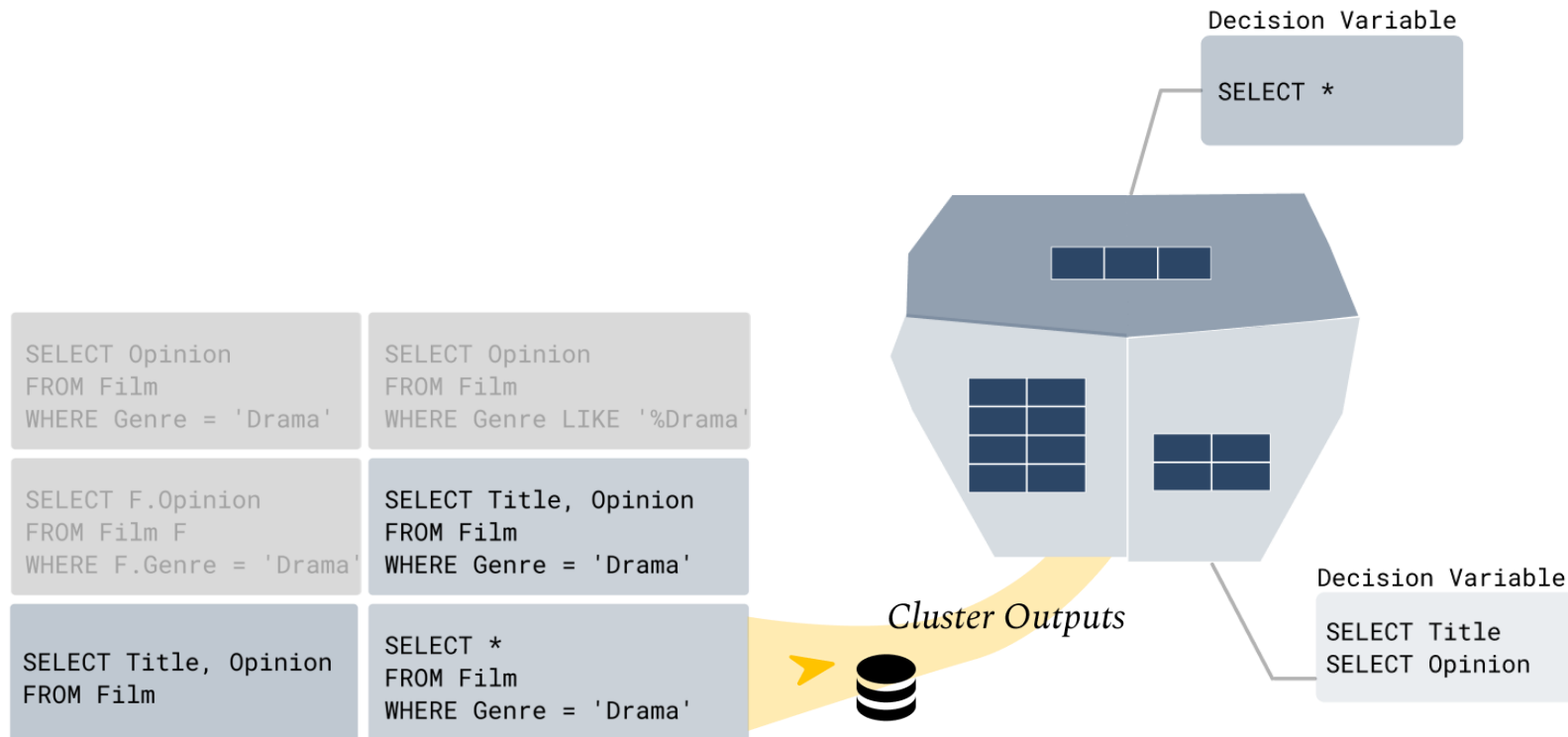
No

5. Filter
Candidates



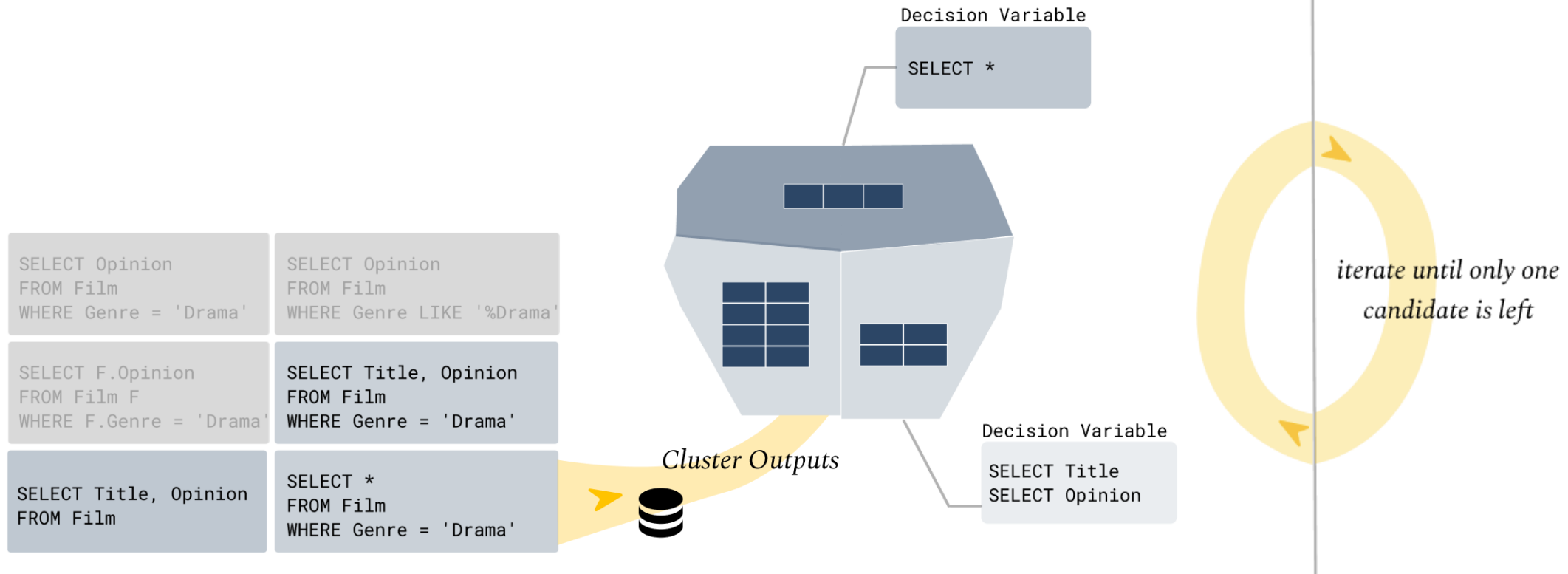
read the paper

Automated Pragmatic Repair



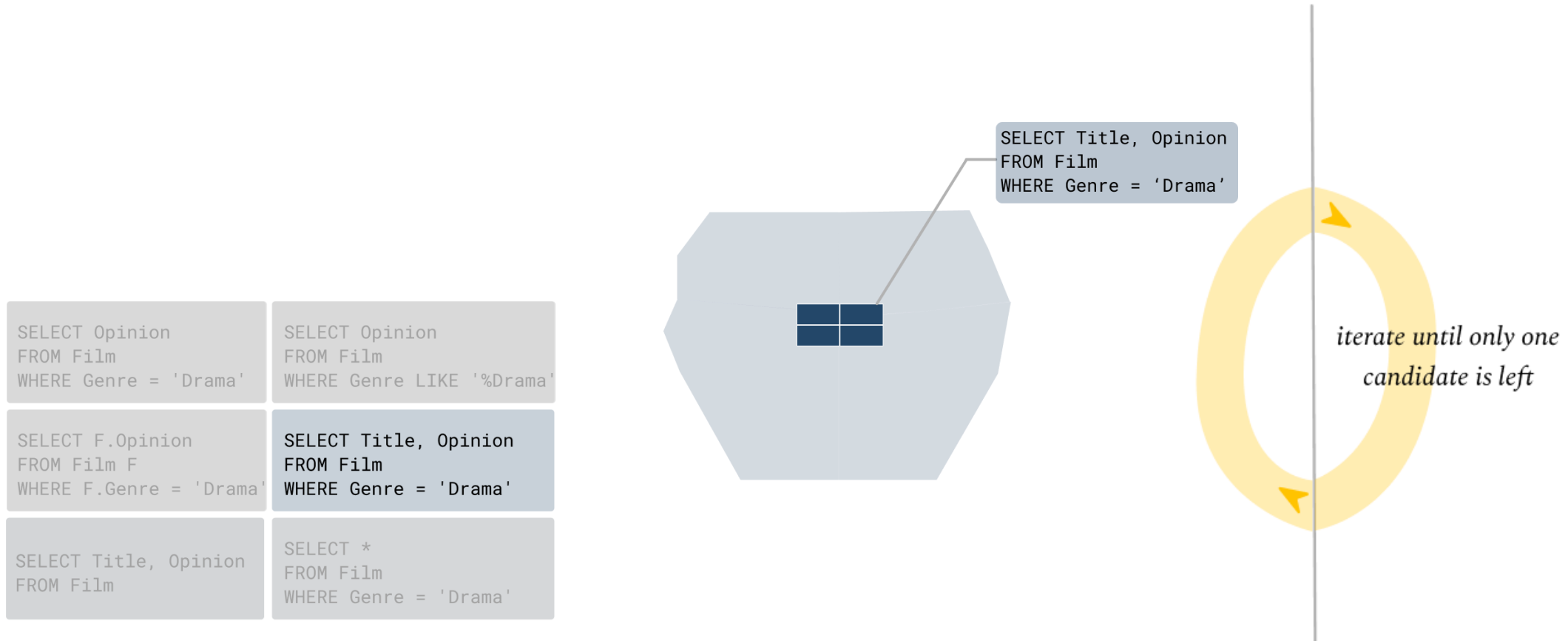


Automated Pragmatic Repair





Automated Pragmatic Repair



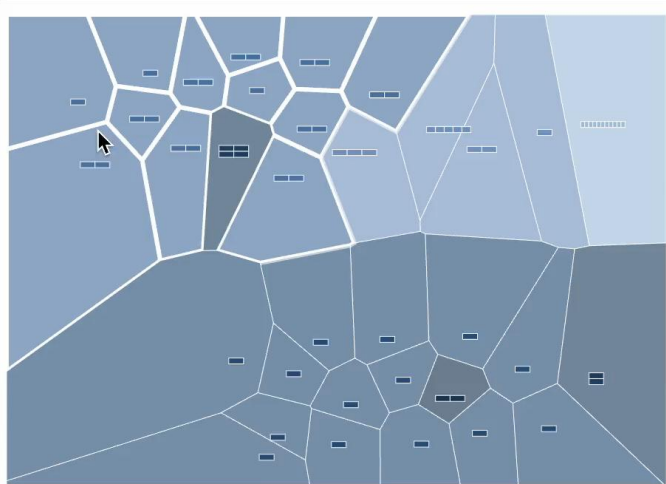


read the paper

Automated Pragmatic Repair

Action Space
Understand Diverse
Interpretations

Action Space





read the paper

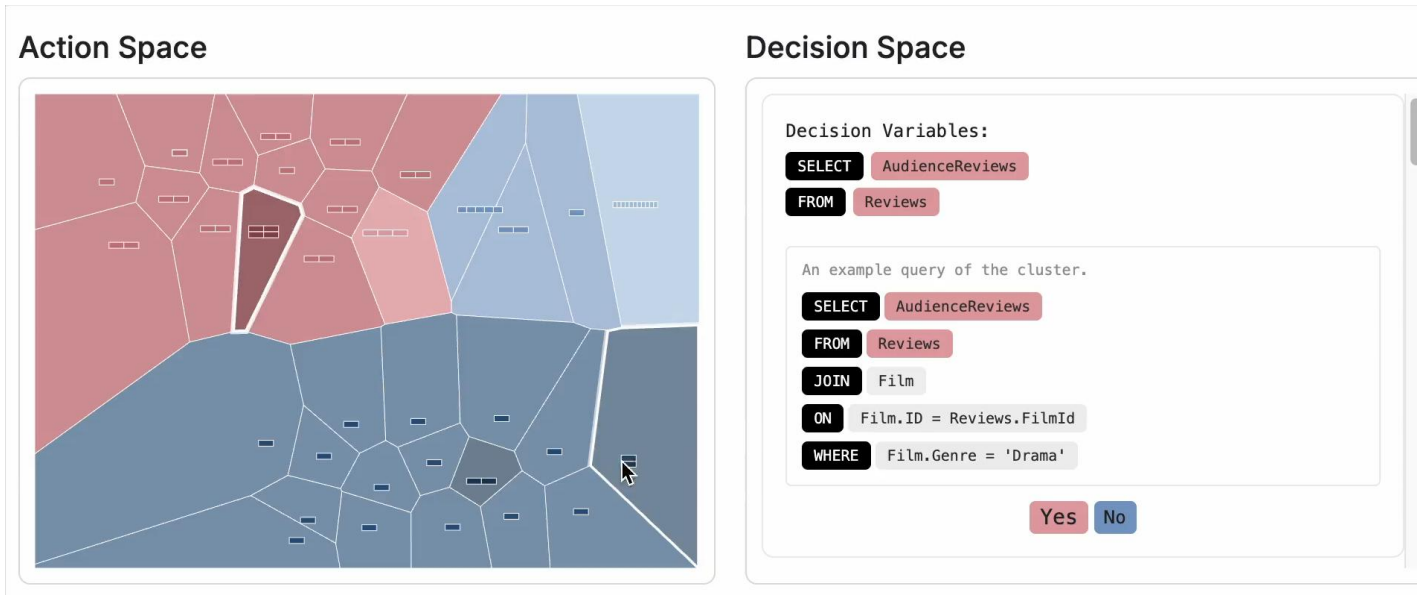
Automated Pragmatic Repair

Action Space

Understand Diverse
Interpretations

Decision Space

Compare Decision Variables
Provide Model Feedback





read the paper

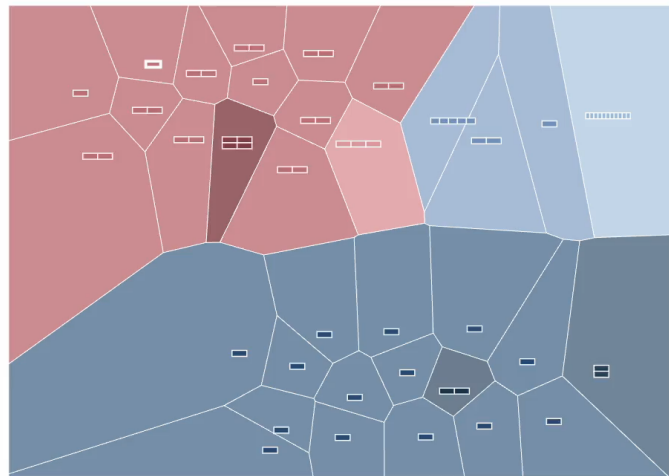
Automated Pragmatic Repair

Action Space
Understand Diverse
Interpretations

Decision Space
Compare Decision Variables
Provide Model Feedback

Predicted Query
Keep Track of
Remaining Uncertainty

Action Space



Decision Space

Decision Variables:

SELECT AudienceReviews
FROM Reviews

An example query of the cluster.

SELECT AudienceReviews
FROM Reviews
JOIN Film
ON Film.ID = Reviews.FilmId
WHERE Film.Genre = 'Drama'

Yes No

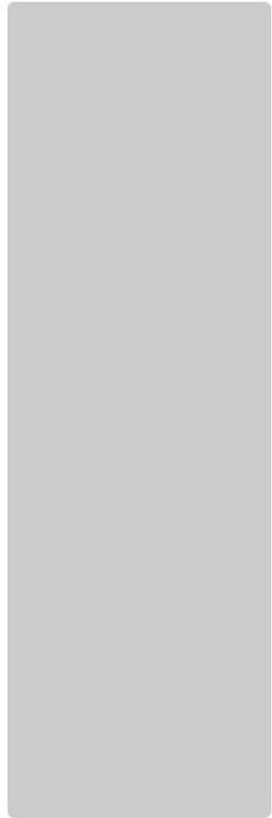
Predicted Query

SELECT Opinion	68%	Yes	No
SELECT AudienceReviews	41%	Yes	No
SELECT *	6%	Yes	No
SELECT CriticName	6%	Yes	No
SELECT Film.Title	3%	Yes	No
SELECT DISTINCT Opinion	15%	Yes	No
SELECT DISTINCT CriticName	3%	Yes	No
FROM Reviews	100%	determined	

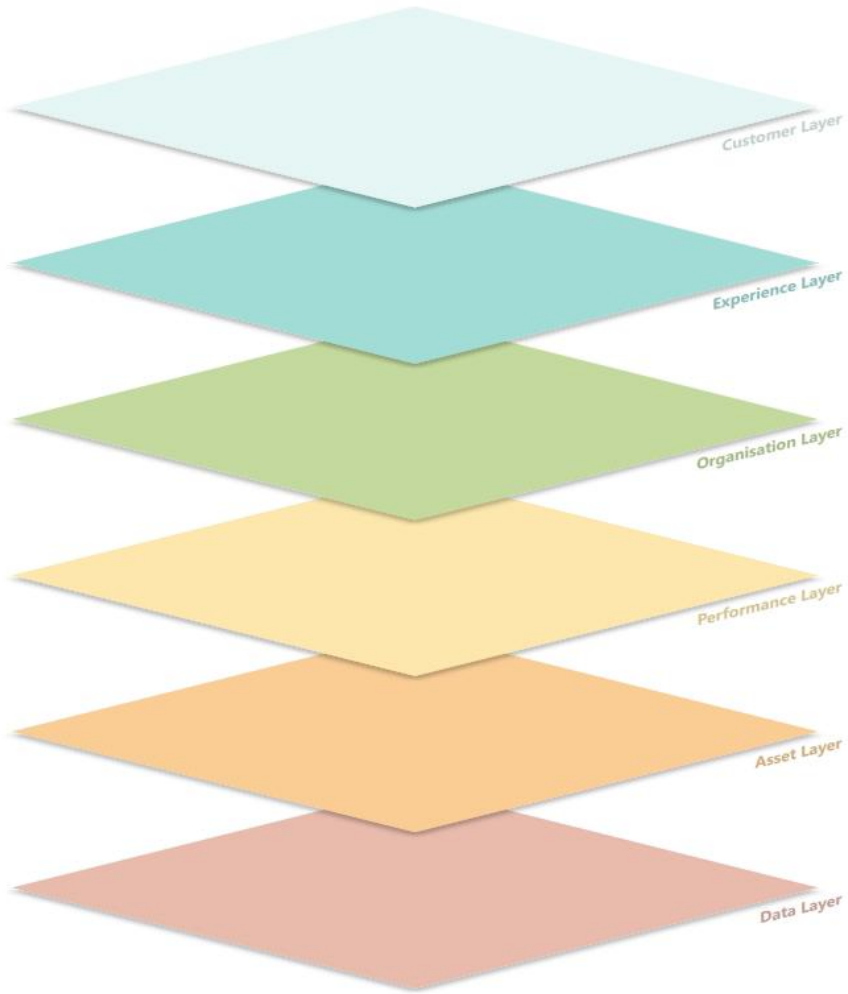
Show Entity Labels: ☒

Enable Debug Mode: ☐

Confidence Threshold:



Unclassified



Recording inactive



Low-Level Contextual Instrumentation



Inspection (Model Explainability)

Probing internal states and latent spaces to understand the "why" behind AI logic.

Steering (Human Feedback)

Pruning wild model growth through semantic interactions and feedback canvases.

Accompaniment (Interaction Timing)

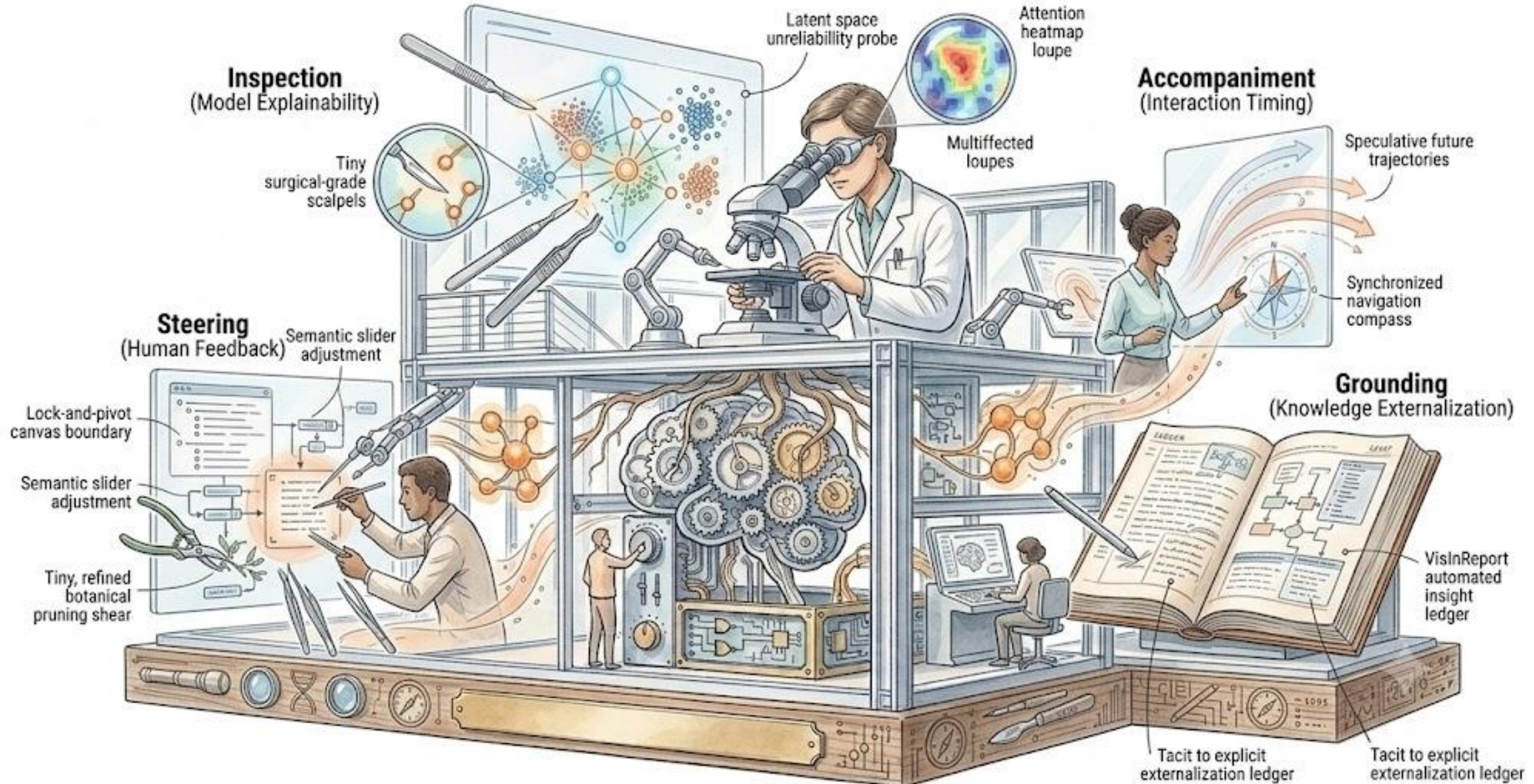
Capturing the sweet spot of interaction timing via adaptive guidance mechanisms.

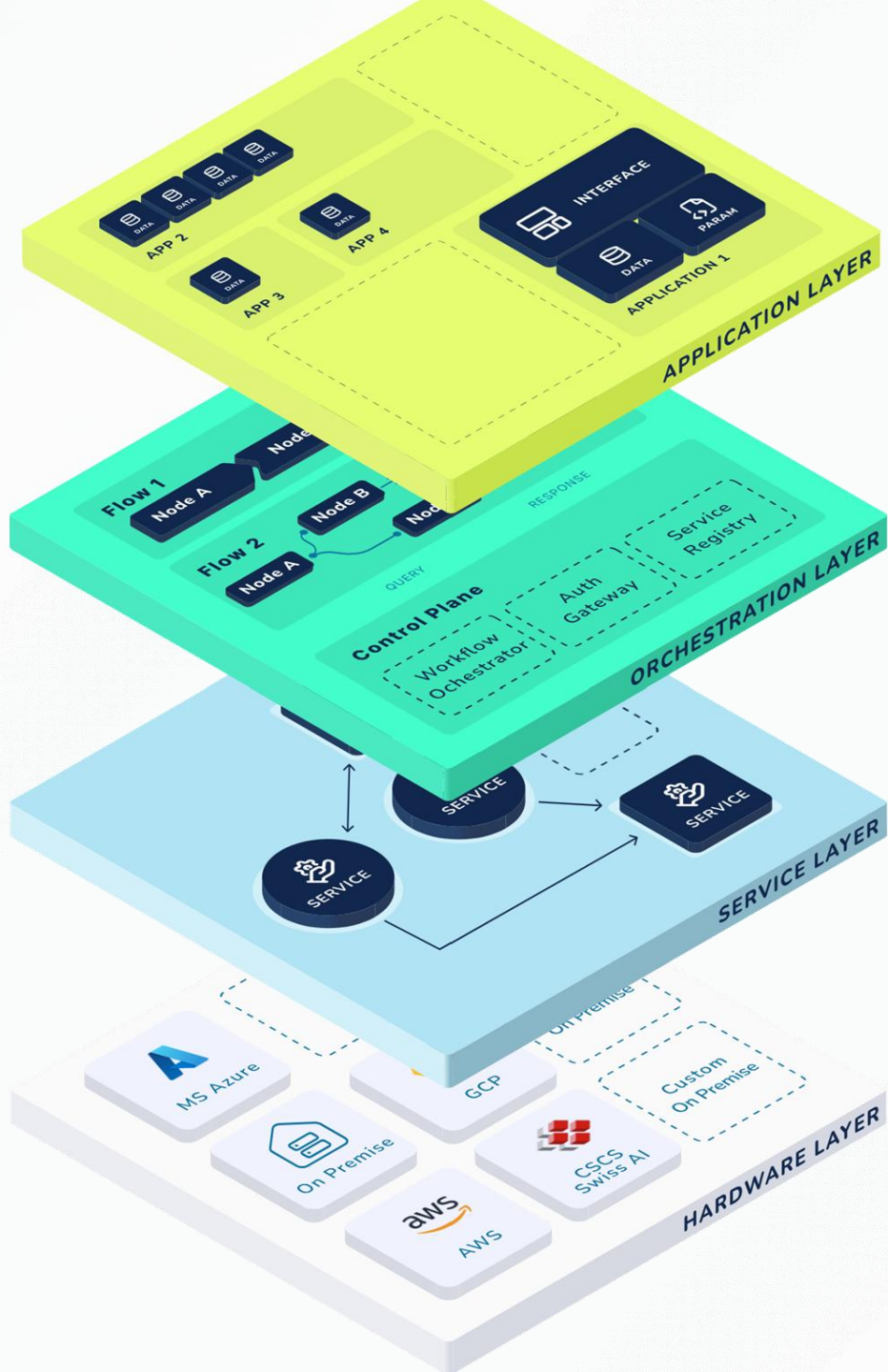
Grounding (Knowledge Externalization)

Externalizing tacit knowledge into explicit, auditable insight ledgers.



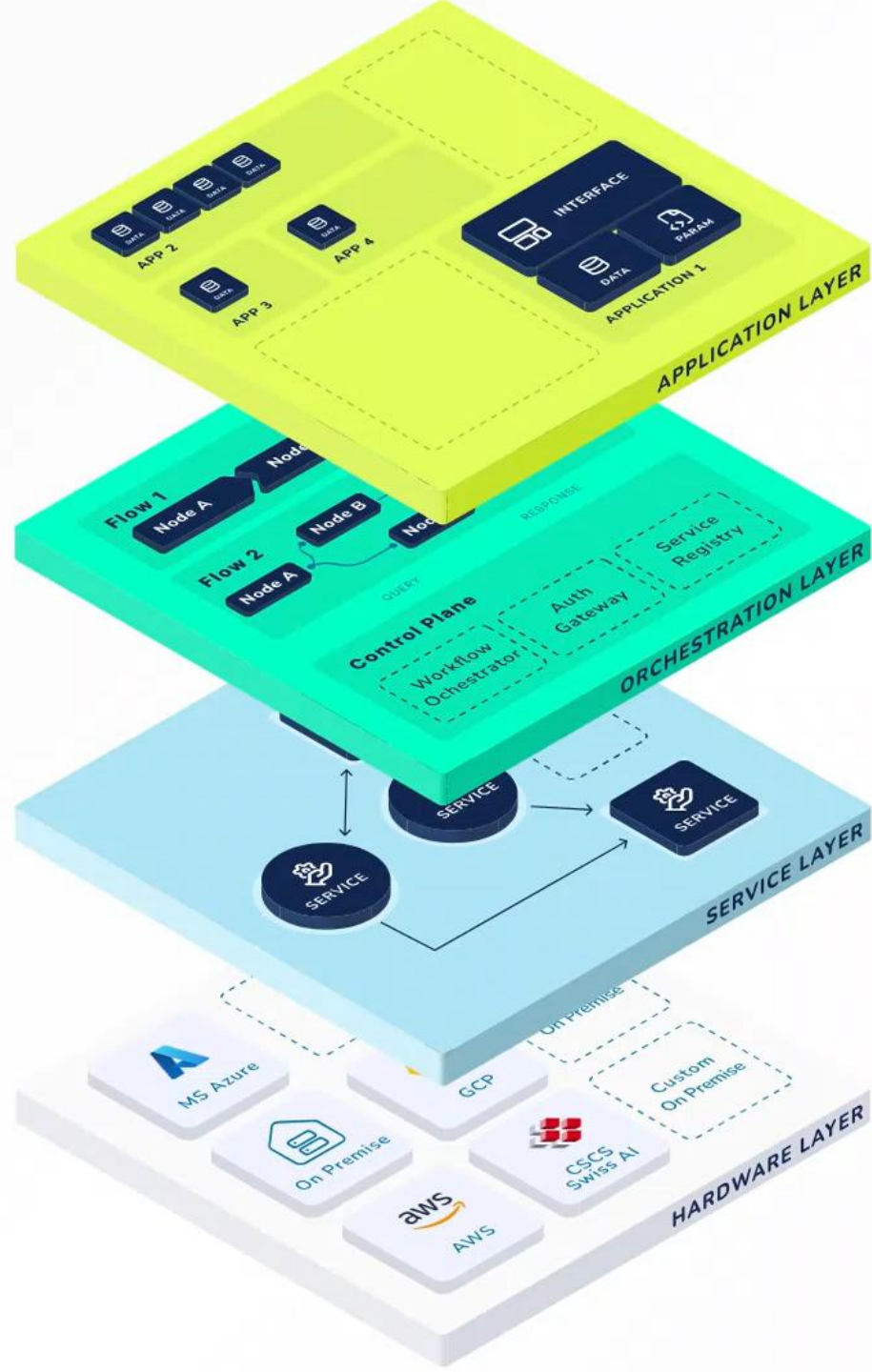
Low-Level Contextual Instrumentation





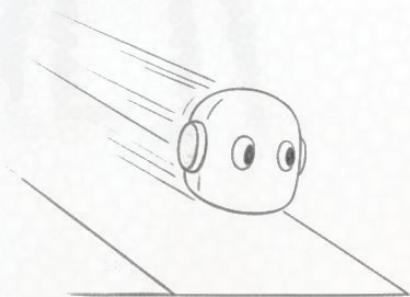
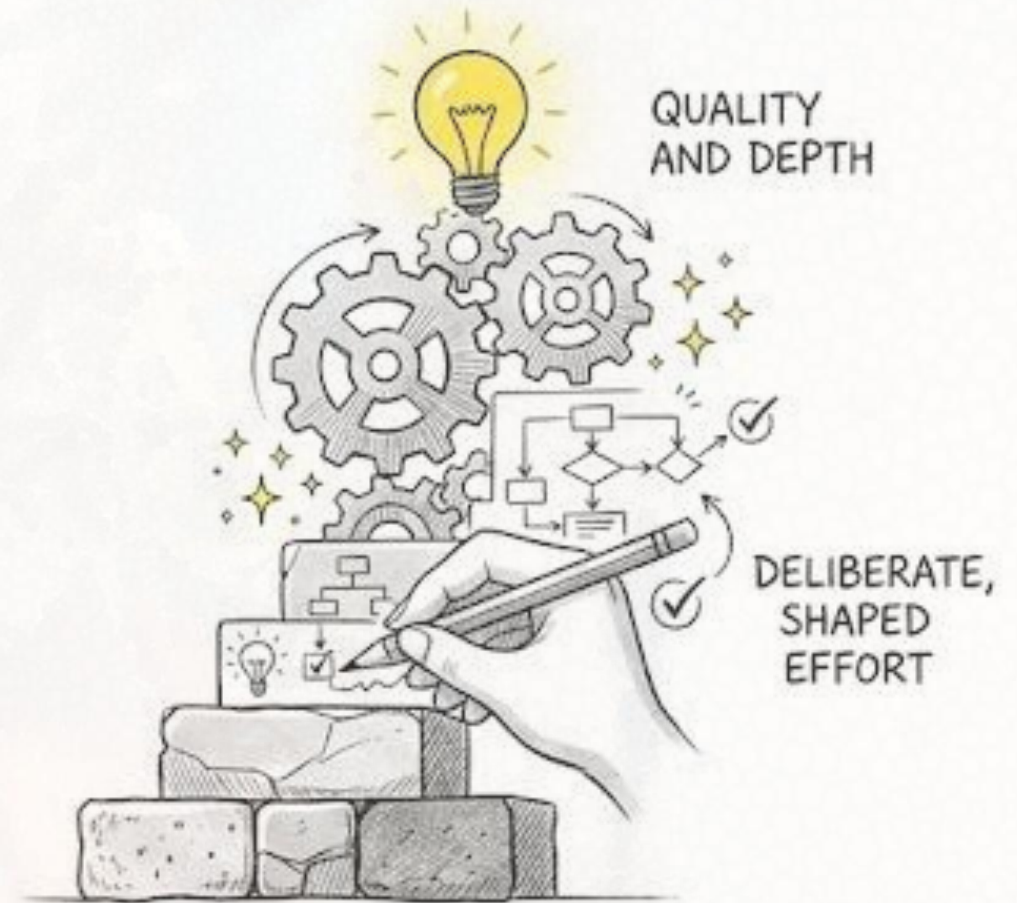
Human-AI Middleware

<https://human-ai.io/>



The IVIA Middleware streamlines the rapid development of interactive AI and data processing workflows. In the following, we will explore its architecture and explain how it helps researchers, developers, and end users to build and configure workflows.

Productive Friction



Our goal is not full automation, because full speed in the wrong direction is **just a faster way to fail.**



**How do we set the right constraints
for the AI's productivity?**

Constrained Co-Creation
Playing in the Bounded Sandbox

THE WORKSPACE

Plan · Design · Monitor · Review

<https://bonsai.ivia.ch/>

A mixed-initiative loop: humans set goals and policies; the orchestrator decomposes work and delegates to specialized units.

Plan

Goals, constraints, brainstorming with feasibility signals, issue board, and configured ADUs.

Design

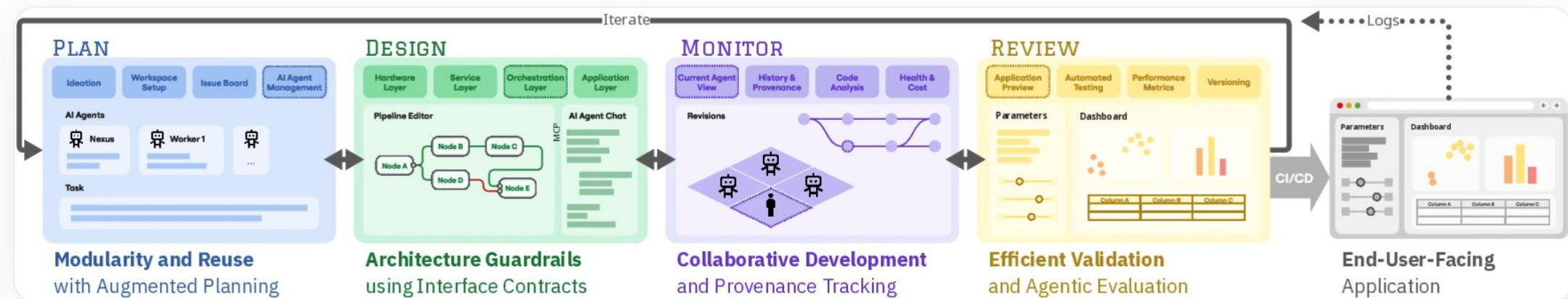
Service registry, typed pipeline editor, application view—new services become nodes without downtime.

Monitor

Agent map plus provenance with semantic zoom—every handoff is observable.

Review

Merge-gated progress, live preview, and issues filed straight from the running prototype.



**Plan****Ready**

Define goals, configure AI Development Units, and set architectural guardrails for your Visual Analytics application.

**Design**

Co-develop across all four architecture layers — from data services and orchestration pipelines to interactive visual interfaces.

**Monitor**

Track provenance, observe human–AI agency handoffs, and maintain supervisory control over the evolving VA system.

**Review**

Validate the co-developed VA application, audit agent contributions, and iterate from visual analysis back to design.

**Ideas**

Scope and brainstorm Visual Analytics application concepts with AI before committing to architecture.



[Call for Action]

designing low-level contextual instruments

Let's study human co-creation
to eliminate human-AI misalignment
and establish human-AI collaboration!

[Call for Action]
**designing low-level
contextual instruments**

Let's study human co-creation
to eliminate human-AI misalignment
and establish human-AI collaboration!



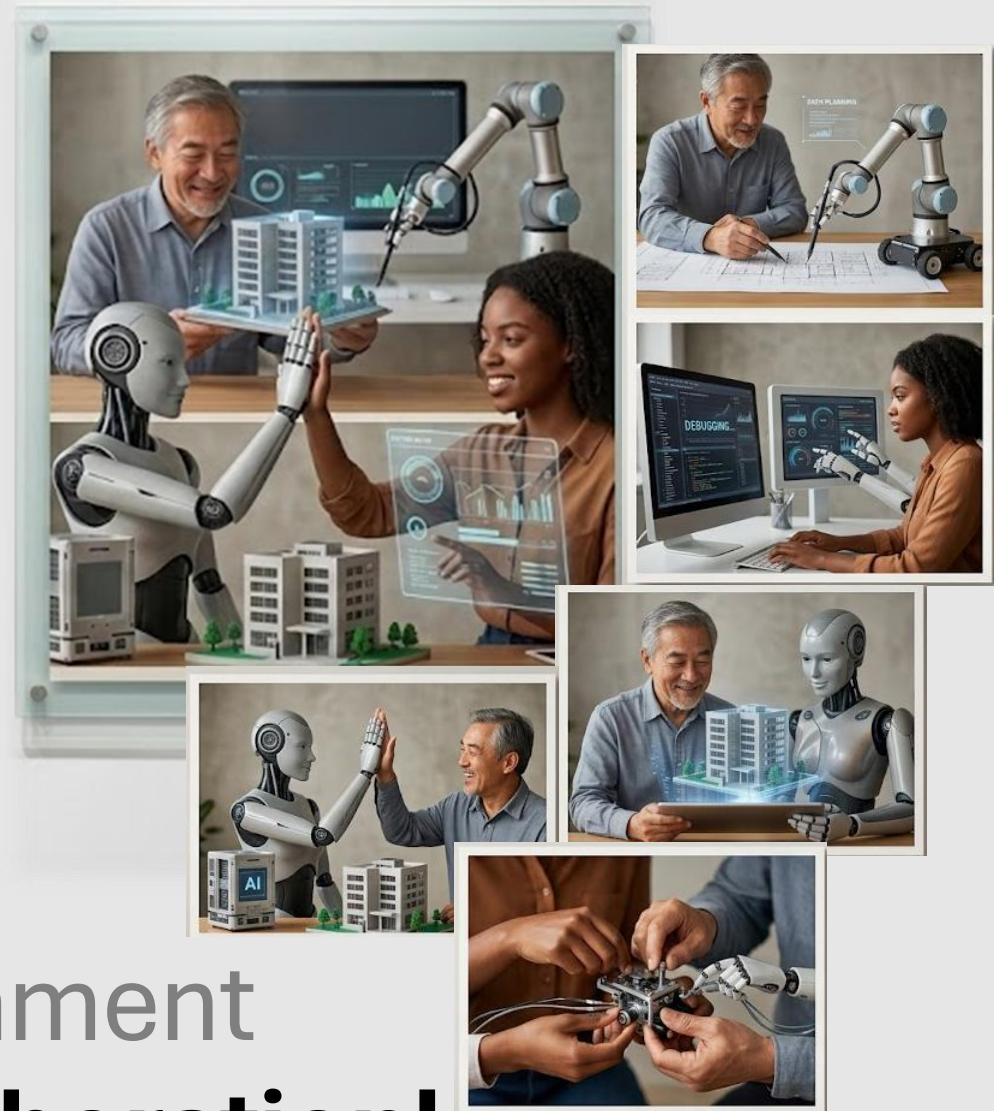
[Call for Action]
**designing low-level
contextual instruments**

Let's study human co-creation
to eliminate human-AI misalignment
and establish human-AI collaboration!



[Call for Action]
**designing low-level
contextual instruments**

Let's study human co-creation
to eliminate human-AI misalignment
and establish human-AI collaboration!



Human-AI Teams can solve more challenging problems than each by themselves!



Our Story...



Research Areas



IVIA Middleware

