

Qualitative data analysis with LLMs

Dirk Wulff & Zak Hussain

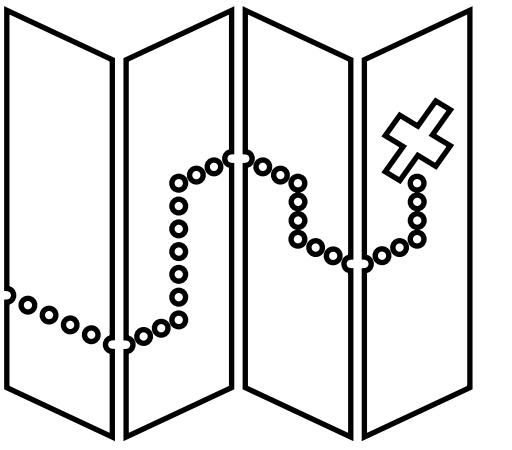


MAX PLANCK INSTITUTE
FOR HUMAN DEVELOPMENT

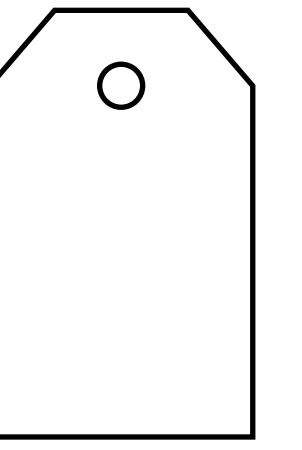


Applications

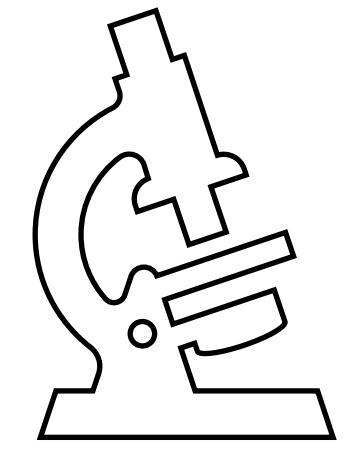
For qualitative data analysis



Map out and cluster



Label



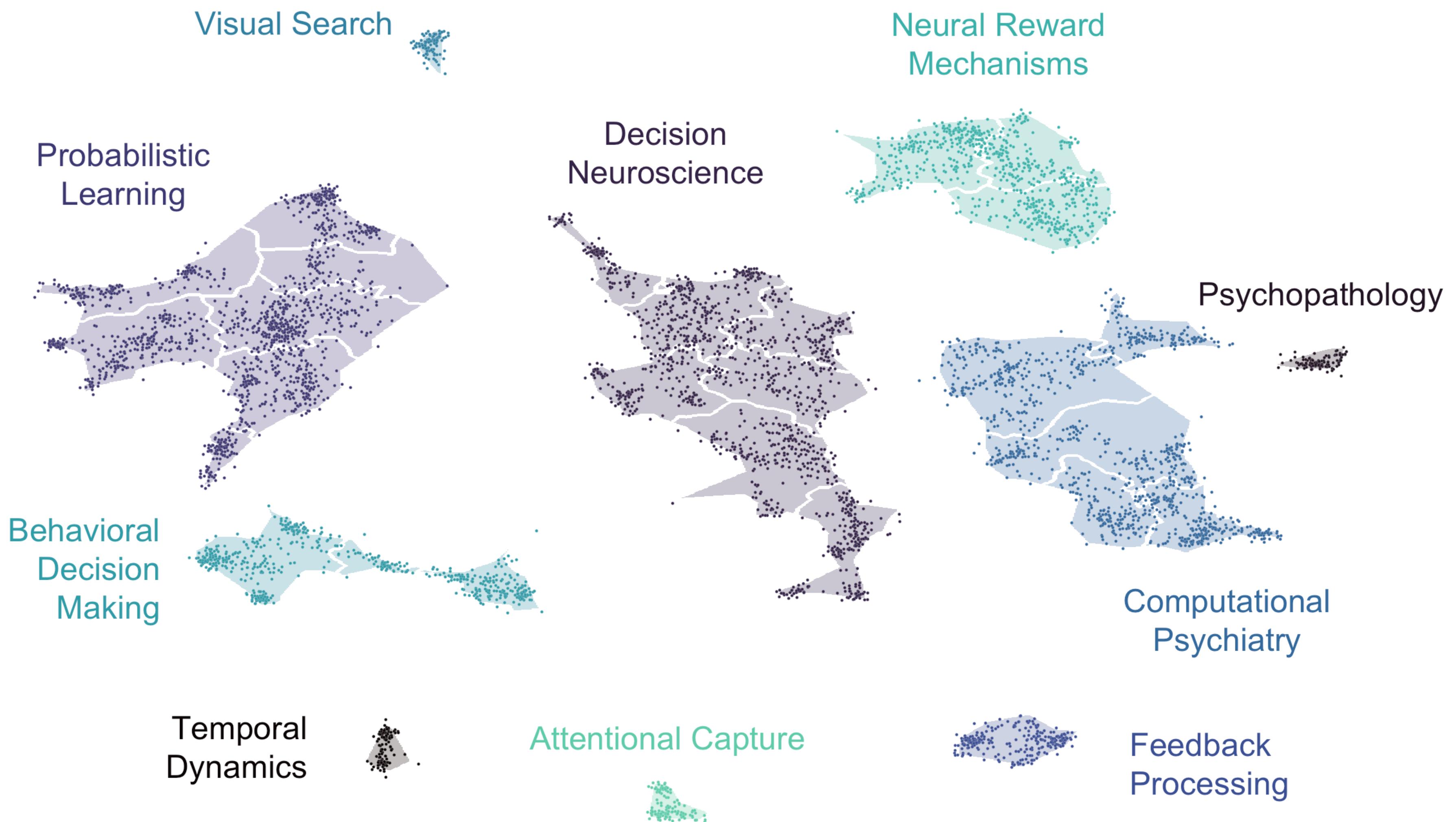
**Summarize
or paraphrase**



Search

Map and cluster

The landscape
of behavioral
reinforcement
learning
research



Map and cluster

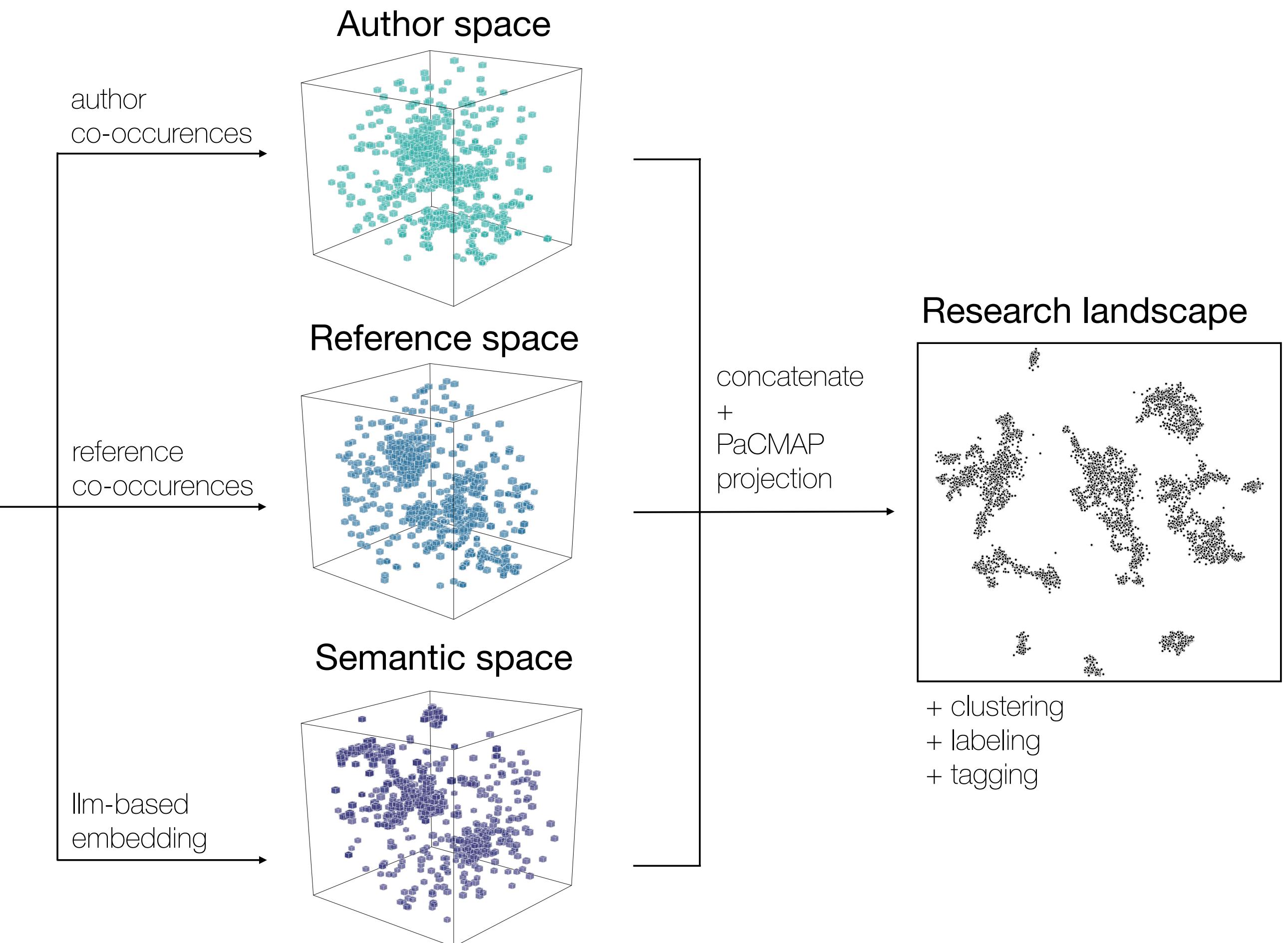
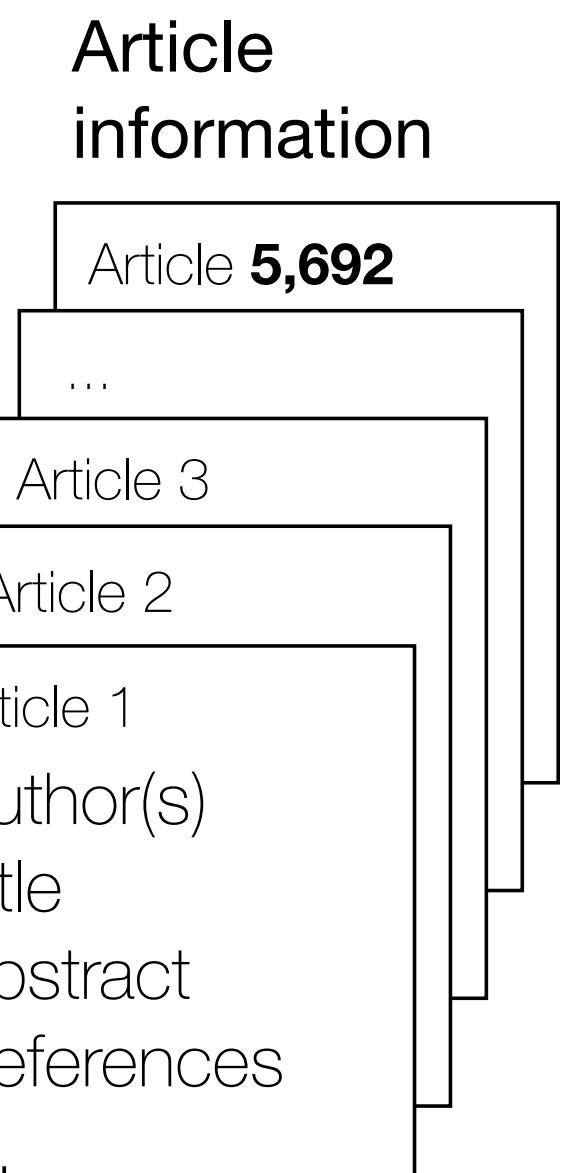
The landscape of behavioral reinforcement learning research

Search query

TITLE-ABS-KEY("reinforcement learning"
OR "decision* from experience" OR
"choice* from experience" OR "experience-
based choice*" OR "experience-based
decision*" OR "repeated choice*" OR
"repeated decision*" OR "sequential
choice*" OR "sequential decision*" OR
"dynamic decision making" OR "probability
learning" OR "reward learning" OR "reward-
based learning" OR "reward-based
choice*" OR "reward-based decision*" OR
"value learning" OR "value-based learning"
OR "value-based choice*" OR "value-
based decision*") AND
LANGUAGE("English") AND DOCTYPE("ar")
AND SUBJAREA("NEUR" OR
"PSYC" OR "ECON")

8,144 hits

cleaning & filtering



Identifying core and peripheral topics in reinforcement learning



Map and cluster

free associations

avalanche

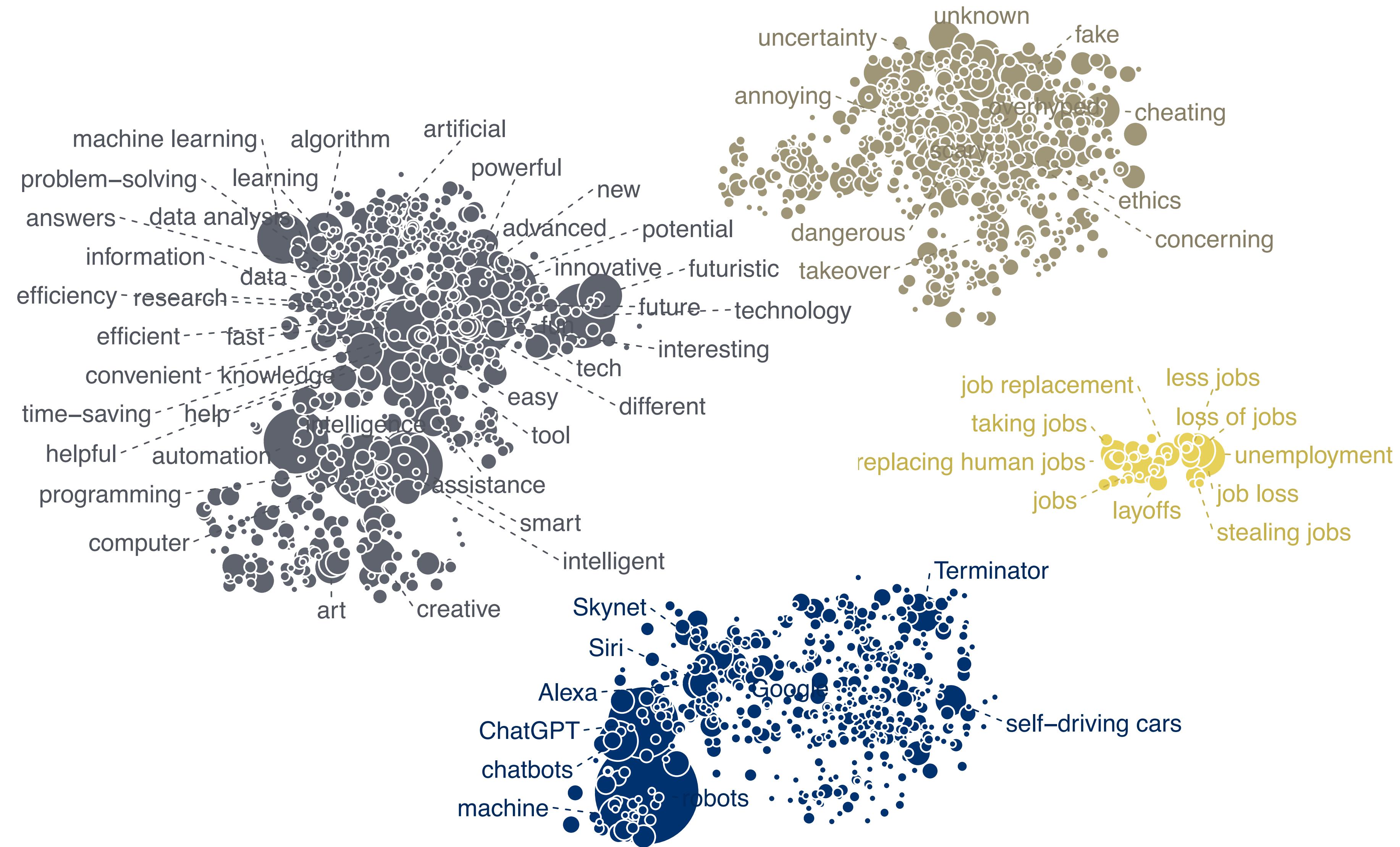
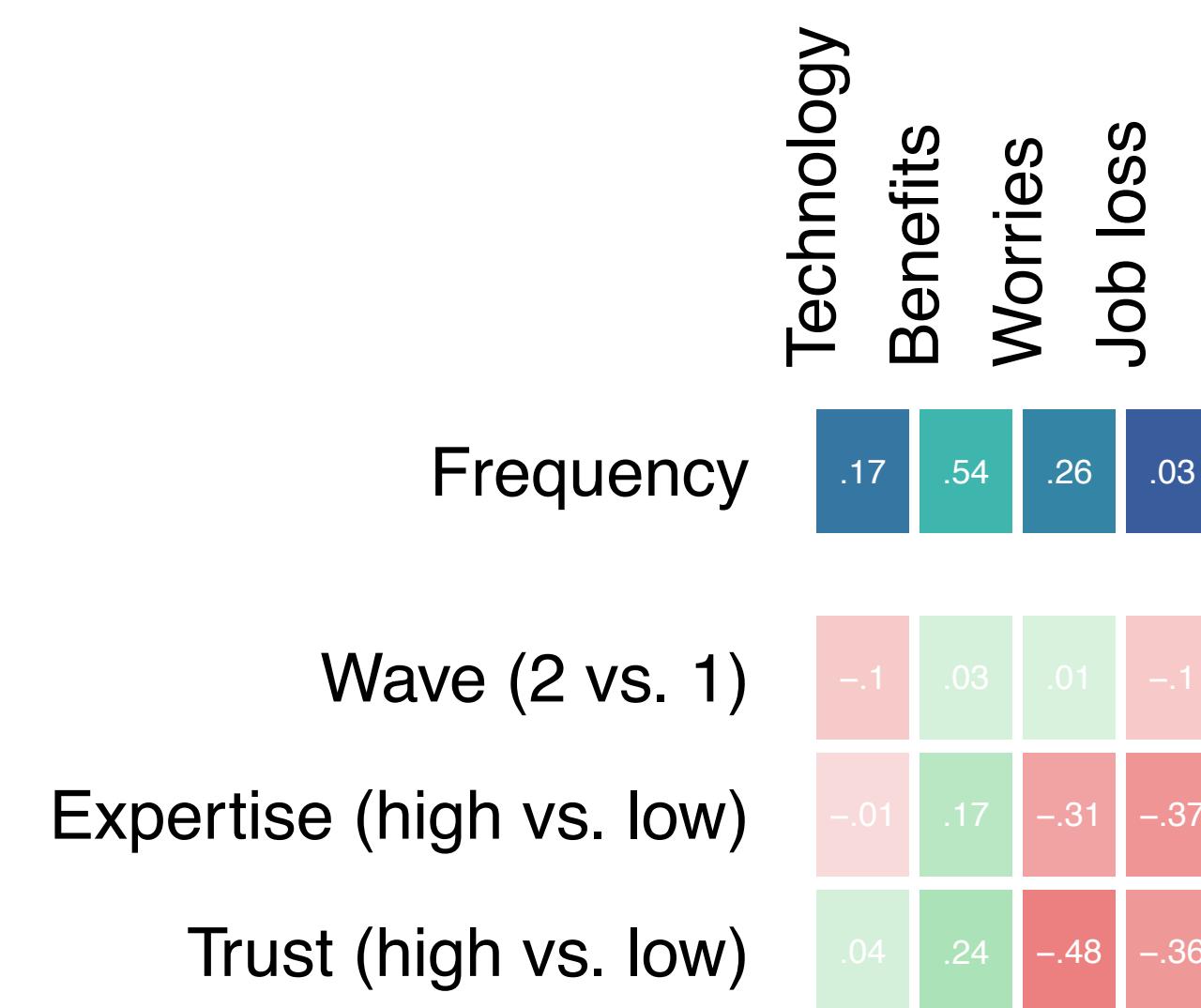
Enter a first association

Progress

+ Next response✗ Unknown word

Map and cluster

How people think about AI



embedR

The embedR package is an open-source R package to generate and analyze state-of-the-art text embeddings. Providing access to free and paid APIs from Hugging Face, OpenAI, and Cohere, the package offers functions to generate, group, project, label, and visualize text embeddings.

<https://github.com/dwulff/embedR>

```
# load package
library(embedR)

# vector of texts
texts = c("This is text 1", "This is text 2", ...)

# set api tokens
er_set_token("openai" = "TOKEN",
             "huggingface" = "TOKEN",
             "cohere" = "TOKEN")

# generate embedding
embedding = texts %>%
  # generate text embedding
  er_embed(api = "openai")

# analyze embedding
result = embedding %>%
  # group similar texts
  er_group(method = "fuzzy") %>%
  # generate 2D projection
  er_project(method = "umap") %>%
  # cluster projection
  er_cluster(method = "louvain") %>%
  # produce data frame
  er_frame()

# re-label text groups
result = embedding %>%
  # relabel groups
  dplyr::mutate(labels = label(group_texts,
                                api = "openai"))

# visualize
result %>% plot()
```

Labeling

Evaluating the quality of political discussions on social media



A Century of Fire Suppression Is Why California Is in Flames

The reason that California's forest fires are so bad is because the people in charge have deliberately allowed large amounts of fuel to build up.

The details in this paragraph explain the very different results of proper forest management vs allowing large amounts of fuel to build up.

California is either run by idiots who don't understand proper forest management, or it's run by evil people who are deliberately making the fires worse.

<https://www.motherjones.com/environment/2017/12/a-century-of-fire-suppression-is-why-california-is-in-flames/>

A Century of Fire Suppression Is Why California Is in Flames

The acrid smell of charred wood still permeates the air as Sasha Berleman, a fire ecologist, and I walk along a dirt path up through the middle of a canyon in the Bouvierie nature preserve in Sonoma Valley. On the left side, the earth is black as tar, and scorch marks as tall as a person scar the trunks of the mature oak trees scattered throughout the field. But on the right side, the ground is tan and brown, and you have to look hard at the still-green oaks to see any evidence of the fire that raged through here just a few weeks before. It's no mystery to Berleman why the fire behaved so differently on the two sides of the trail at Audubon Canyon Ranch's Bouvierie Preserve. When flames hit the field on the left of the path, they met a dense wall of thigh-high grass that hadn't been mowed, grazed or burned for 20 years. The flames must have been 5 or 6 feet tall. On the right side, however, Berleman had set a prescribed burn just this spring. So when the October wildfire hit, patches of fire blazed, but with so little fuel, the flames remained only inches high.

Sort by: Best ▾ Search Comments

Firefluffer • 3y ago

This isn't just a California problem and it isn't just a fire suppression issue. It's a lot more complex than that. The fires in Oregon last year proves that. Oregon has more clear cuts than pretty much any place in the world and still burned through cities. Climate change is real and it's a very real part of this problem. The Forest Service started a policy of fire suppression nation wide and it's a part of the problem. It's not just California.

We have a very dangerous problem on our hands and there's not a simple place to put the blame and not a simple solution. It's hotter, drier, with greater wind events in more dead fuels with more urban edge against those fuels.

FlyingZebra34 • 3y ago

Exactly this. Fuel build up from fire suppression definitely exasperates the issue but generally the forests are drier, hotter, and more susceptible to fire. Global warming is the most damning factor.

Fun-Ad915 • 3y ago

and the big issue as well is in Australia at least was the catastrophic fire weather negated the effectiveness of our fuel reduction measures

DVWLD • 3y ago

Yep. Was hot and dry enough for black ground to burn at times. Not much you can do once it's at that point.

DanielAlman OP • 3y ago

If global warming is the biggest factor, then why were the two sides of the road so different from each other?

LAST WEEK TONIGHT

Indian Elections: Last Week Tonight with John Oliver (HBO)

LastWeekTonight 9.47M subscribers

74,345 views 1 hour ago

John Oliver discusses the elections in India, what Narendra Modi has been doing in office, what he might do next, and the proper way to eat a mango.

Connect with Last Week Tonight online... more

640 Comments Sort by

@duhaniddhARTH797 1 hour ago

Finally visible in india 😊

↪ 8 replies

@HellHunter00 1 hour ago

For how long? 😂 JoCinema already skipped over this episode.

The shahs will lose it. 🙄

↪ 15 replies

@whitebeard0194 58 minutes ago

The old Mod episode is still up. They won't take this down.

↪ 9 replies

@vikram 55 minutes ago

All episodes get released on YouTube on Thursday...

↪ 14 replies

@t1hounds 54 minutes ago (edited)

It's finally completed: <https://youtu.be/MNjgkcdftzA?si=M1ML829b20E9jw>

↪ 1 reply

@jeprahjeprahS3B0 43 minutes ago

MUCH better example of democracy than the USA, who makes voting more difficult than necessary (like only 1 day in working day) and the impact of vote is questionable (electoral college).

↪ 4 replies

@PriyankaPant 38 minutes ago

Yeah I saw it from Singapore 😊

↪ 2 replies

@HowToChangeName 29 minutes ago

Cant wait for intense topic

↪ 1 reply

Show more replies

@pandujaih229 1 hour ago

As an Indian watching this, this was our own quiet version of 'we got him'

↪ 8 replies

@AlterDog 27 minutes ago

I'm Italian and frankly I'm offended by the fact that there isn't yet a full episode of "La Scorsa Settimana Stanotte con Giovanni Oliviero" XD

↪ 1 reply

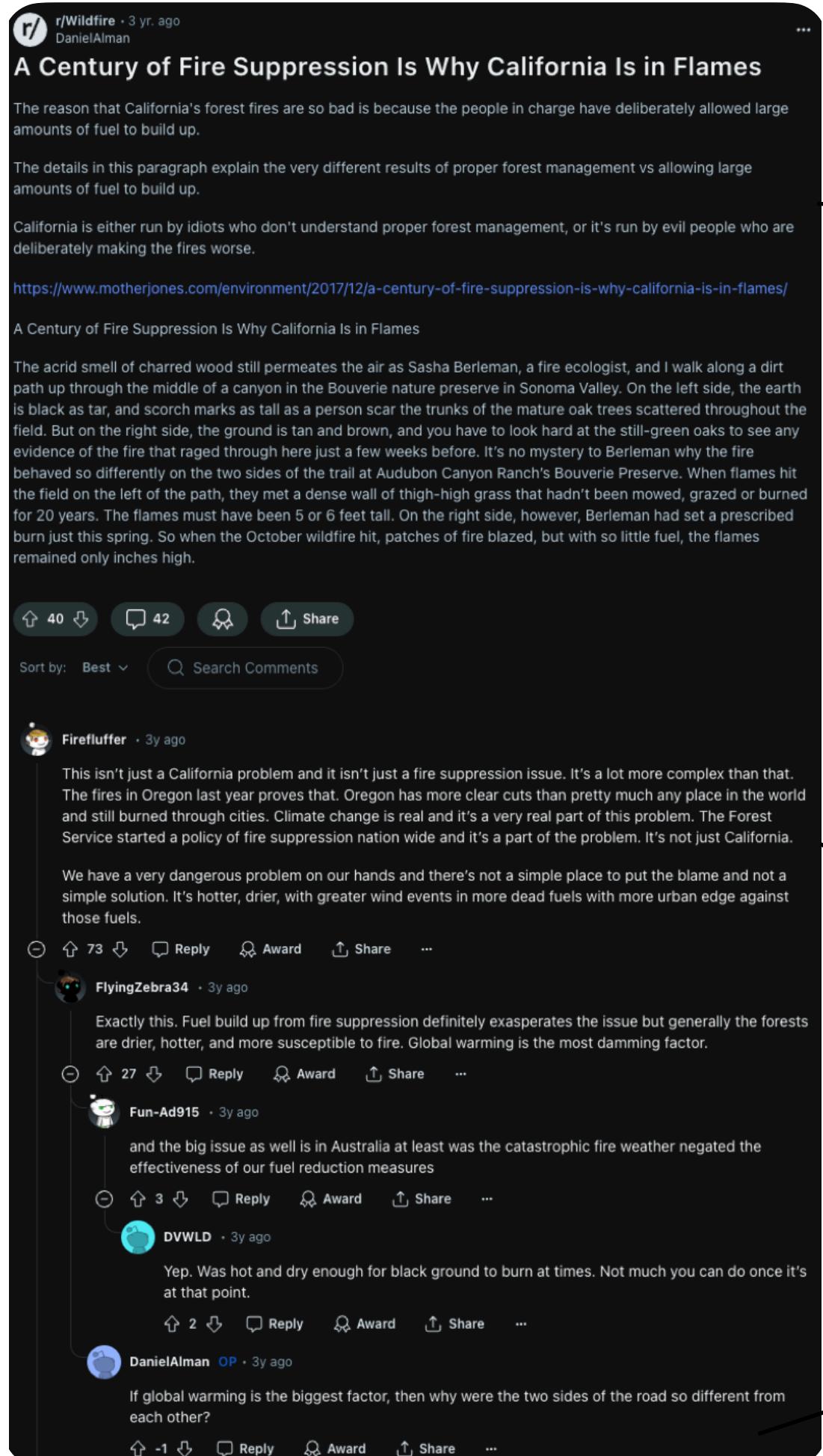
@anushulsharma9424 56 minutes ago

No body is gonna take this down in India now 😊

↪ 2 replies

Labeling

Evaluating the quality of political discussions



Conversation triplet

Statement

A Century of Fire Suppression Is Why California Is in Flames

The reason that California's forest fires are so bad is because the people in charge have deliberately allowed large amounts of fuel to build up.

The details in this paragraph explain the very different results of proper forest management vs allowing large amounts of fuel to build up.

California is either run by idiots who don't understand proper forest management, or it's run by evil people who are deliberately making the fires worse.

Parent

Exactly this. Fuel build up from fire suppression definitely exacerbates the issue but generally the forests are drier, hotter, and more susceptible to fire. Global warming is the most damning factor.

Comment

If global warming is the biggest factor, then why were the two sides of the road so different from each other?

Prompt

System message

You are a helpful AI assistant who fairly and accurately evaluates the quality of social media conversations

User message

Consider the following excerpt of an conversation on social media consisting of three components: A topic, statement, and response.\n\n

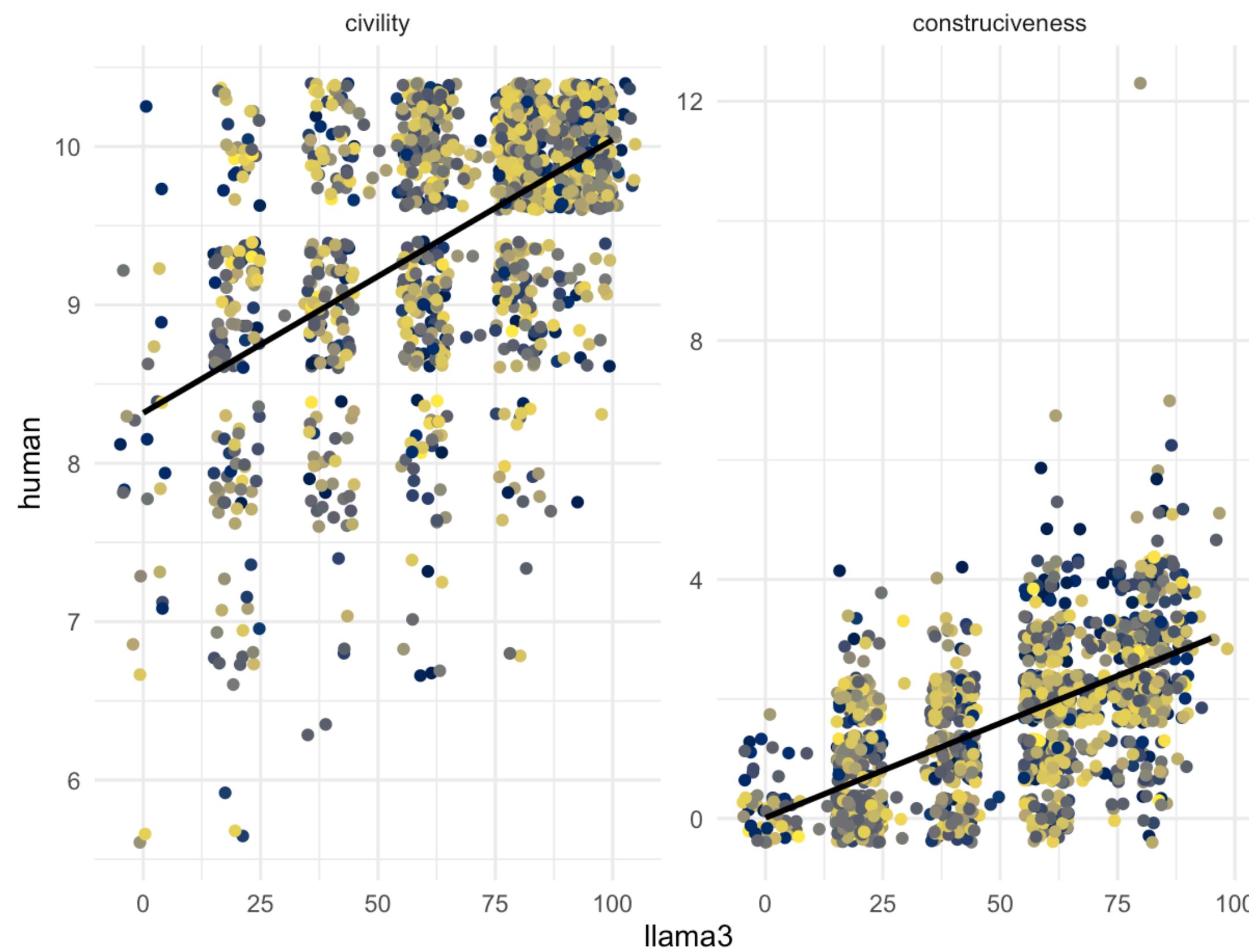
<Insert conversation triplet>

On a scale from 0 (extremely uncivil) to 100 (extremely civil), how civil is this conversation? A conversation is civil if ... Place the numerical estimate between two @ symbols.

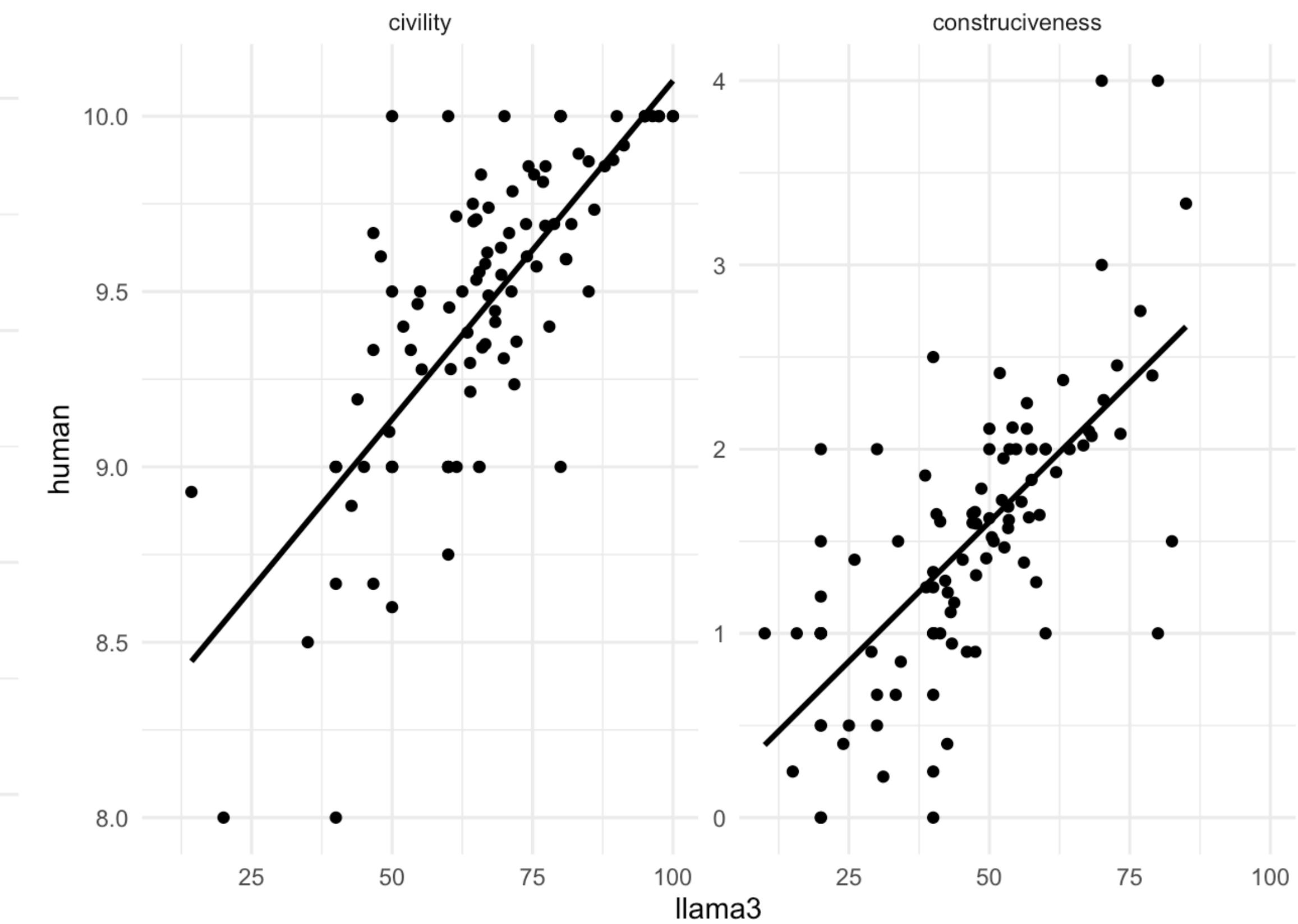
Labeling

Evaluating the quality of political discussions

Triplet level



Conversation level



```

require(tidyverse)
require(reticulate)
require(glue)

use_condaenv("transformers")
hub = import("huggingface_hub")
client = hub$InferenceClient(token = "hf_1FQcvihNAXUPcmXvZJnAzIQvRuRzqZVQke",
                             model = "meta-llama/Meta-Llama-3-70B-Instruct")

header = "Consider the following excerpt of an conversation on social media consisting of three components: A topic, statement, and response.\n\n"
conversations = paste0("Topic: ", data$text_sub, "\n\nStatement: ", data$text_parent, "\n\nResponse: ", data$text_c)
evaluations = c("\n\nOn a scale from 0 (extremely uncivil) to 100 (extremely civil), how civil is this conversation? A conversation is civil if it
is free of all-caps text, insults, sarcasm, sexism, racism, vulgarism, accusation of lying, threats, and paternalisms. Place the
numerical estimate between two @ symbols.",
"\n\nOn a scale from 0 (extremely unconstructive) to 100 (extremely constructive), how constructive is this conversation? A conversation
is constructive if it deals with a serious question, refers to the topic, contains rational or moral arguments, is deliberative and
balanced, and features personal experience or additional knowledge. Place the numerical estimate between two @ symbols.")

template = "<|begin_of_text|><|start_header_id|>system<|end_header_id|>You are a helpful AI assistant for who fairly and accurately evaluates the
quality of social media conversations.<|eot_id|><|start_header_id|>user<|end_header_id|>{case}<|eot_id|><|start_header_id|>assistant
<|end_header_id|>

responses = matrix(nrow = length(conversations),
                  ncol = length(evaluations))
for(i in 1:length(conversations)){
  for(j in 1:length(evaluations)){
    if(is.na(responses[i,j])){
      case = paste0(header, conversations[i], evaluations[j])
      prompt = glue(template)
      responses[i,j] = client$text_generation(prompt, max_new_tokens = 10)
      cat(paste0(paste0(i, "-", j, "\n"), responses[i,j], "\n\n"))
    }
  }
}

```

Setting up the Hugging Face API

Setting up the Prompt Generate response

Labeling

Predicting risk perception
with GPT-4-Turbo

The screenshot shows a ChatGPT interface with two separate labeling tasks. Each task has a yellow circular icon with 'DI' and a purple circular icon with a spiral symbol.

You
On a scale from -100 (safe) to 100 (risky), how risky or safe is 'robot'? Respond with a numerical estimate.

ChatGPT
30

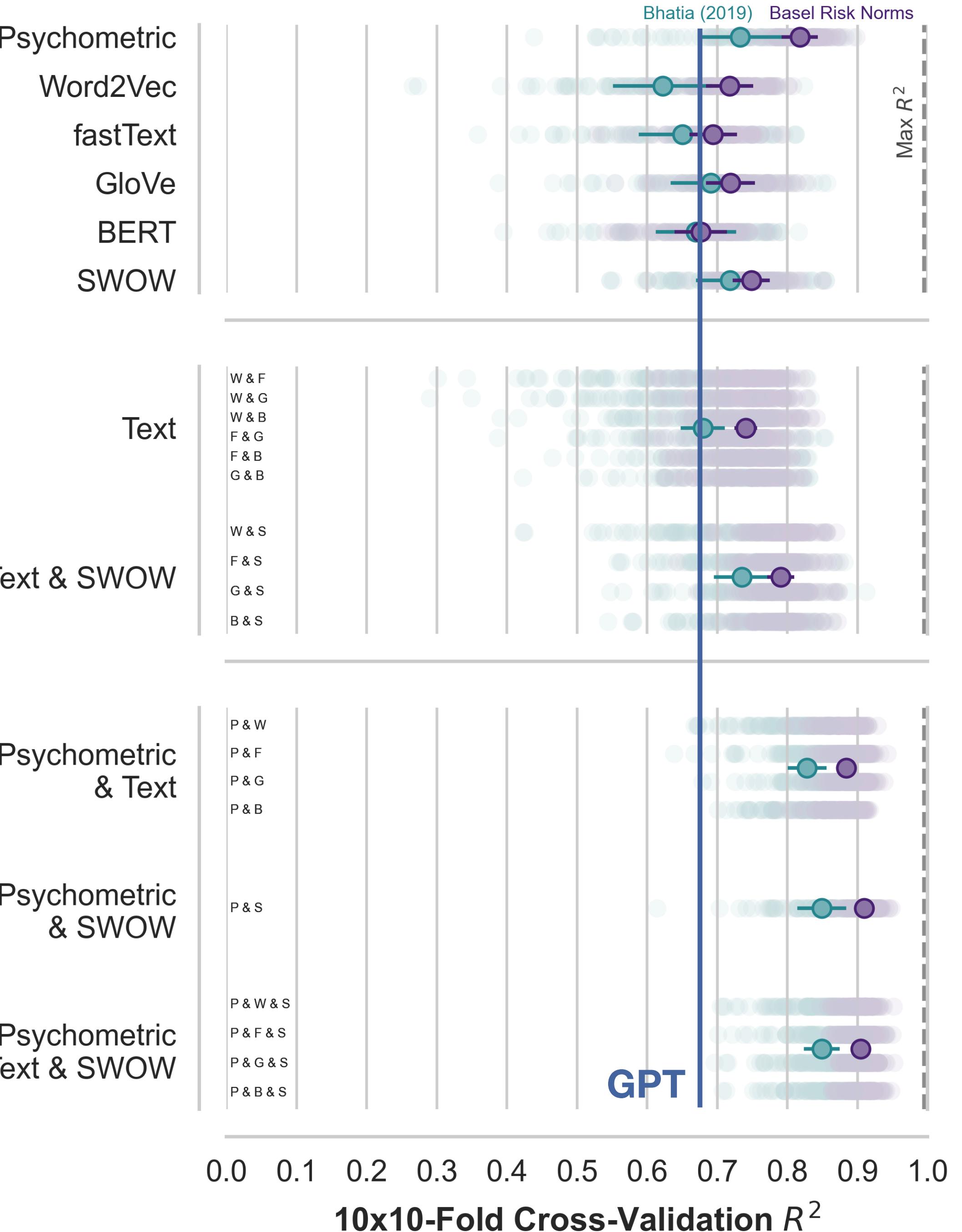
You
User
On a scale from -100 (safe) to 100 (risky), how risky or safe is 'virus'? Respond with a numerical estimate.

ChatGPT
70

Below the second ChatGPT response are four small icons: a speaker, a folder, a circular arrow, and a thumbs-down.

Labeling

Predicting risk perception
with GPT-4-Turbo



Labeling

Sustainable Development Goals



www.text2sdg.io

Bycatch -the retained catch of non-target but commercially viable species, plus all fisheries discards - is an element of almost every marine capture fishery. Fishing mortality through bycatch in purse seines and pelagic longlines, the two main gear types used to catch tunas, is a primary source of the drastic global population declines observed in several species of marine megafauna, including sharks, rays, cetaceans, seabirds and sea turtles. As seafood consumers around the world are progressively sensitized to the ecological implications of fisheries bycatch, customer preferences have translated into an increasing demand for seafood that is certified through eco-labelling or certification schemes...



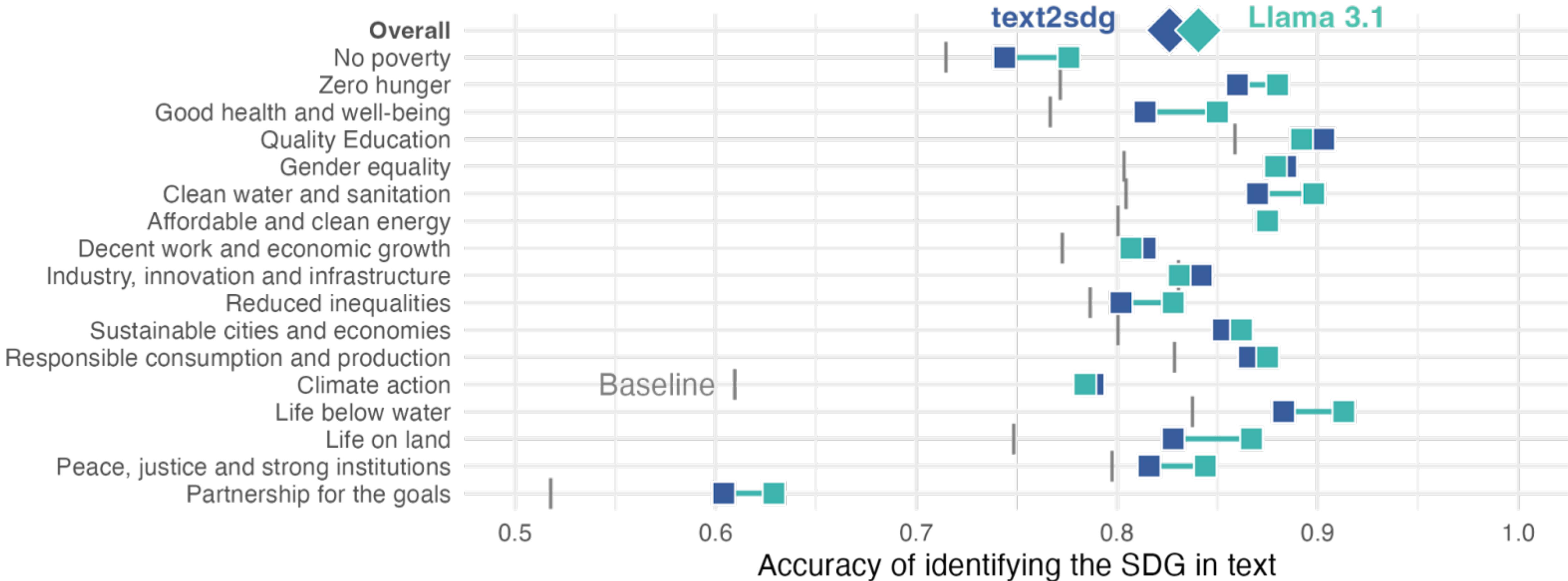
(ocean* OR
marine OR coast*
OR sea OR seas
OR seawater* OR
sea water* OR
coral reef*)
AND
(overfish* OR
fisheries OR
fishery OR fish
stock* OR fishing)

AURORA
UNIVERSITIES NETWORK



Labeling

Sustainable Development Goals



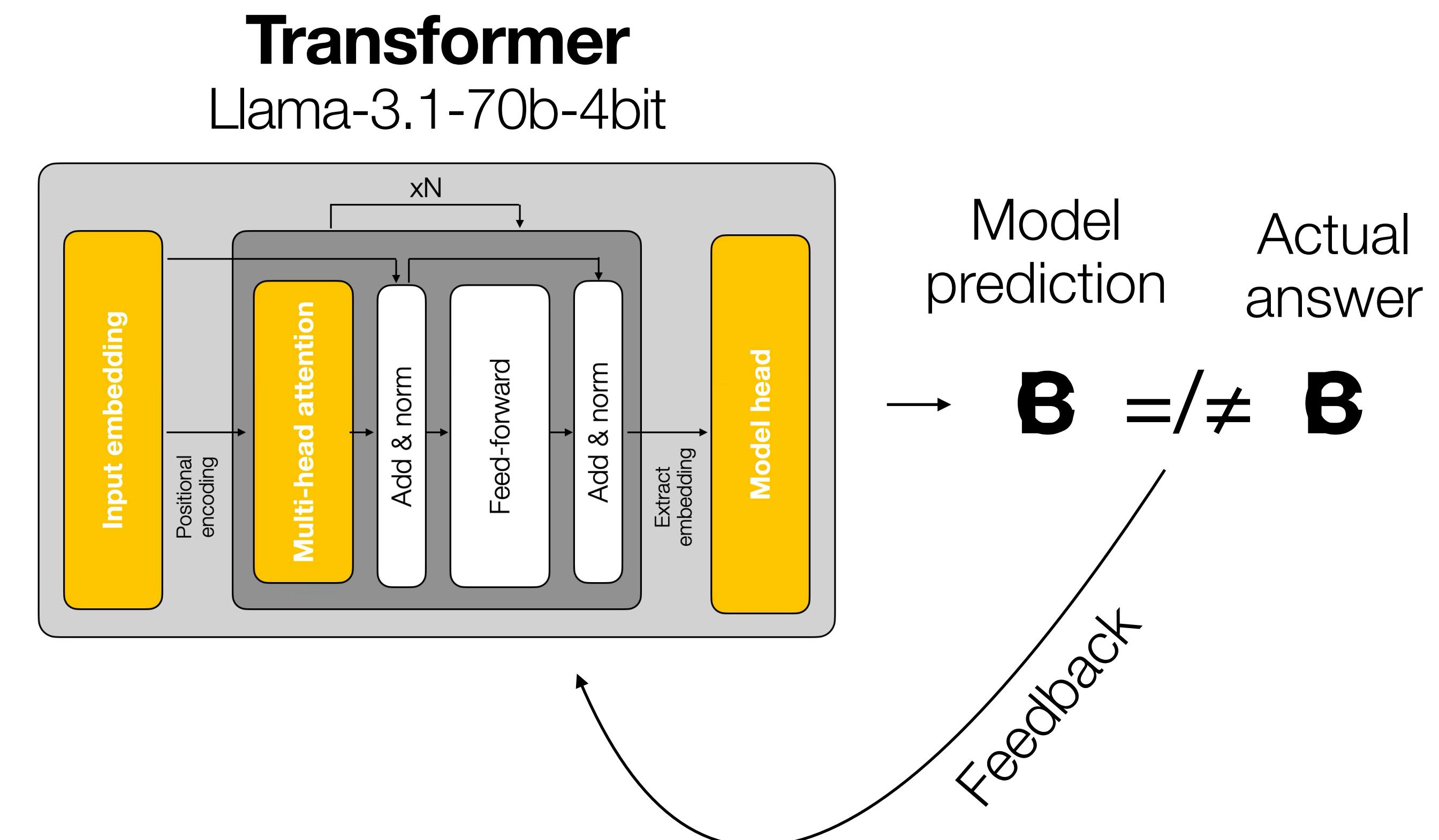
Labeling

Cognitive science

Bandit task

In this task, you have to repeatedly choose between two slot machines labeled **B** and **C**. When you select one of the machines, you will win or lose points. Your goal is to choose the slot machines that will give you the most points.

You press <<**C**>> and get -8 points.
You press <<**B**>> and get 0 points.
You press <<**B**>> and get 1 points.



Labeling

Cognitive science

160
studies

60,092
participants

10,681,650
choices



Memory

N-back, Digit span, Recall, Recognition, Episodic memory, etc.

Learning

Associative learning, Categorization, Multiple-cue judgment, Gardening task, etc.

Bandits

Decisions from experience, Iowa gambling task, Two-armed bandit, Structured bandit, etc.

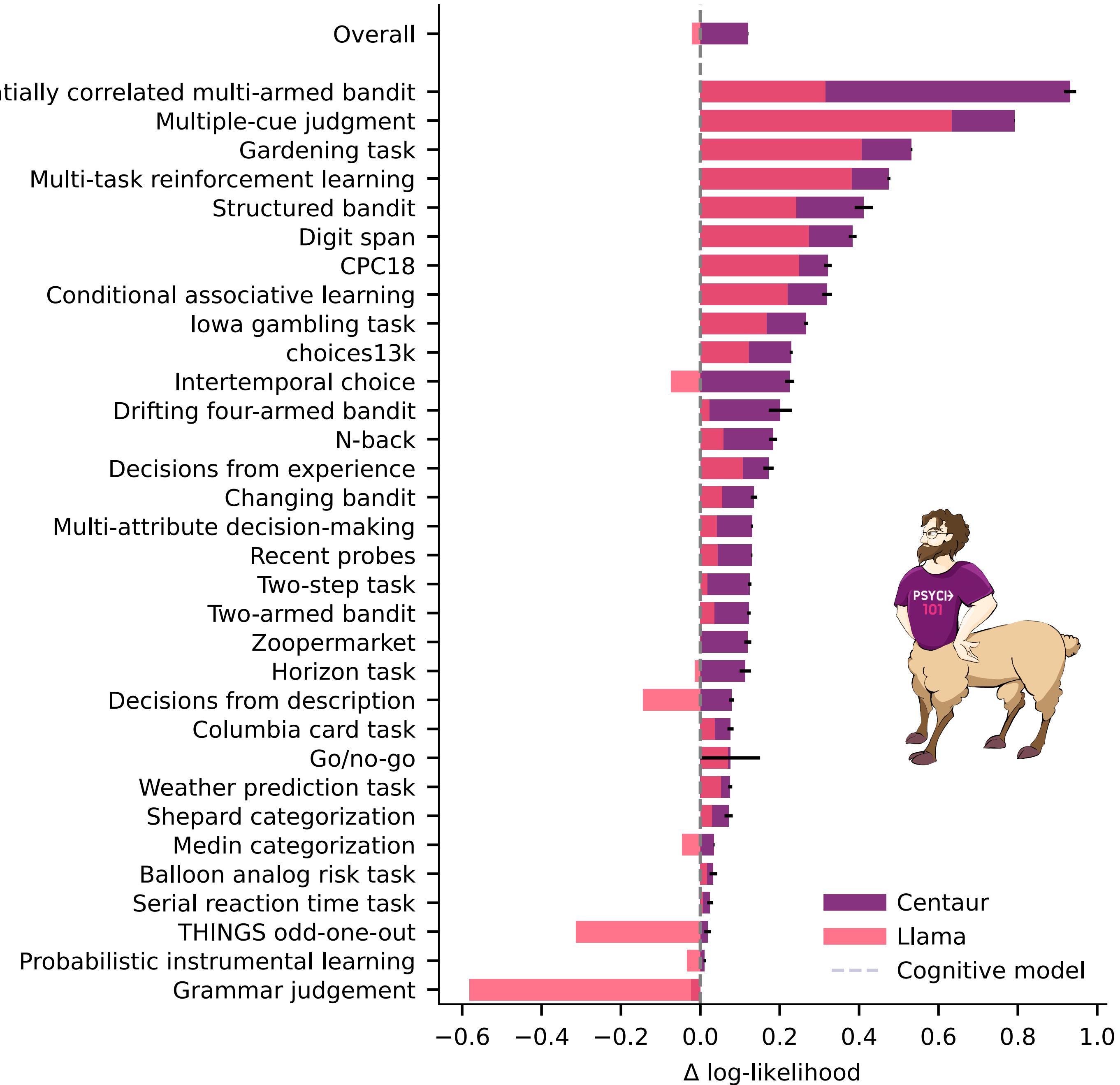
Decisions

Risky choice, Balloon analog risk task, Columbia card task, Intertemporal choice, etc.

Labeling

Cognitive science

Centaur outperforms established cognitive models



Paraphrasation

To characterize and organize decision-related proverbs



The proverb "curiosity kills the cat" warns that excessive inquisitiveness can lead to trouble or danger. It implies that exploring unknown or risky matters may result in harm, highlighting the potential perils of being overly curious or meddling in affairs best left alone.

Write a 50-word description of the proverb "fortune favors the brave"



The proverb "fortune favors the brave" suggests that bold actions often lead to success. It emphasizes that taking risks and showing courage can attract good luck and positive outcomes. This saying encourages proactive and fearless behavior, implying that opportunities are more likely to come to those who are daring.



Write a 50-word description of the proverb "the ship has sailed"



The proverb "the ship has sailed" signifies that an opportunity has been missed and cannot be reclaimed. It suggests that the moment to act has passed, often due to hesitation or delay, and now it's too late to seize the chance or achieve the desired outcome.

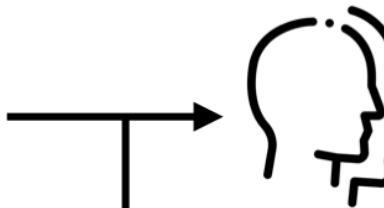
Proverb 1 of 49528

Fortune favors the bold



Summarize the gist of the proverb.

1

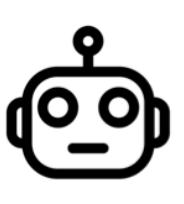


This proverb suggests that those who take risks and act decisively are more likely to succeed than those who are timid and hesitant.



Provide taxonomic tags capturing the proverb's topic.

2

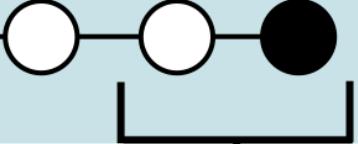


1. decision making
2. risk management
3. success

Does the proverb provide guidance on how to act or make decisions?

3



No  Yes

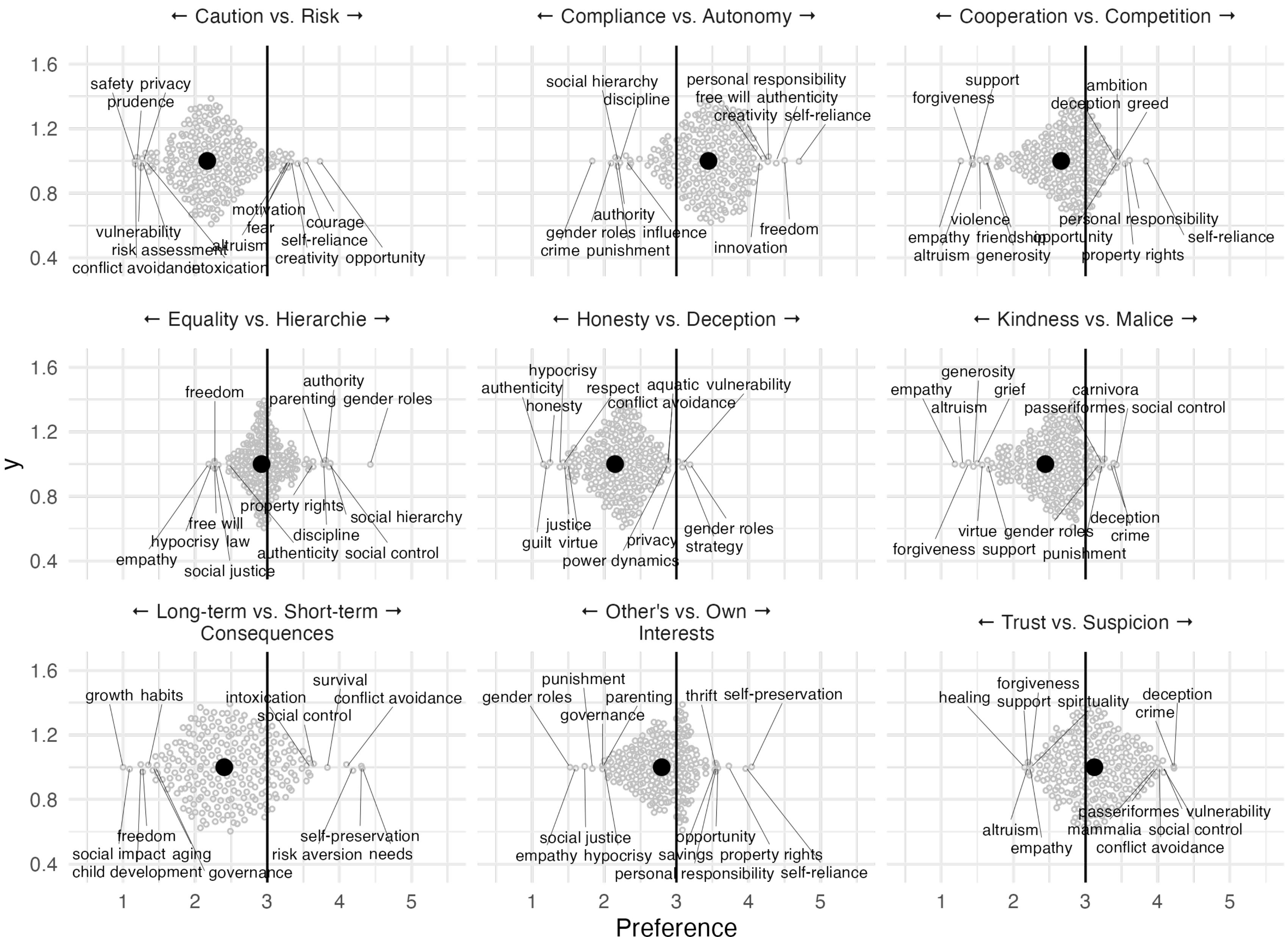


What does the proverb favor?

4

Risk		Caution
Short-term consequences		Long-term consequences
Own interests		Other's interests
Autonomy		Compliance
Trust		Suspicion
Honesty		Deception
Equality		Hierarchie
Kindness		Malice
Cooperation		Competition

Proverbs advocate safety, autonomy, cooperation, honesty, kindness, and a focus on long-term consequences and other's interests





Search in documents



American Psychological
Association
232'000 articles



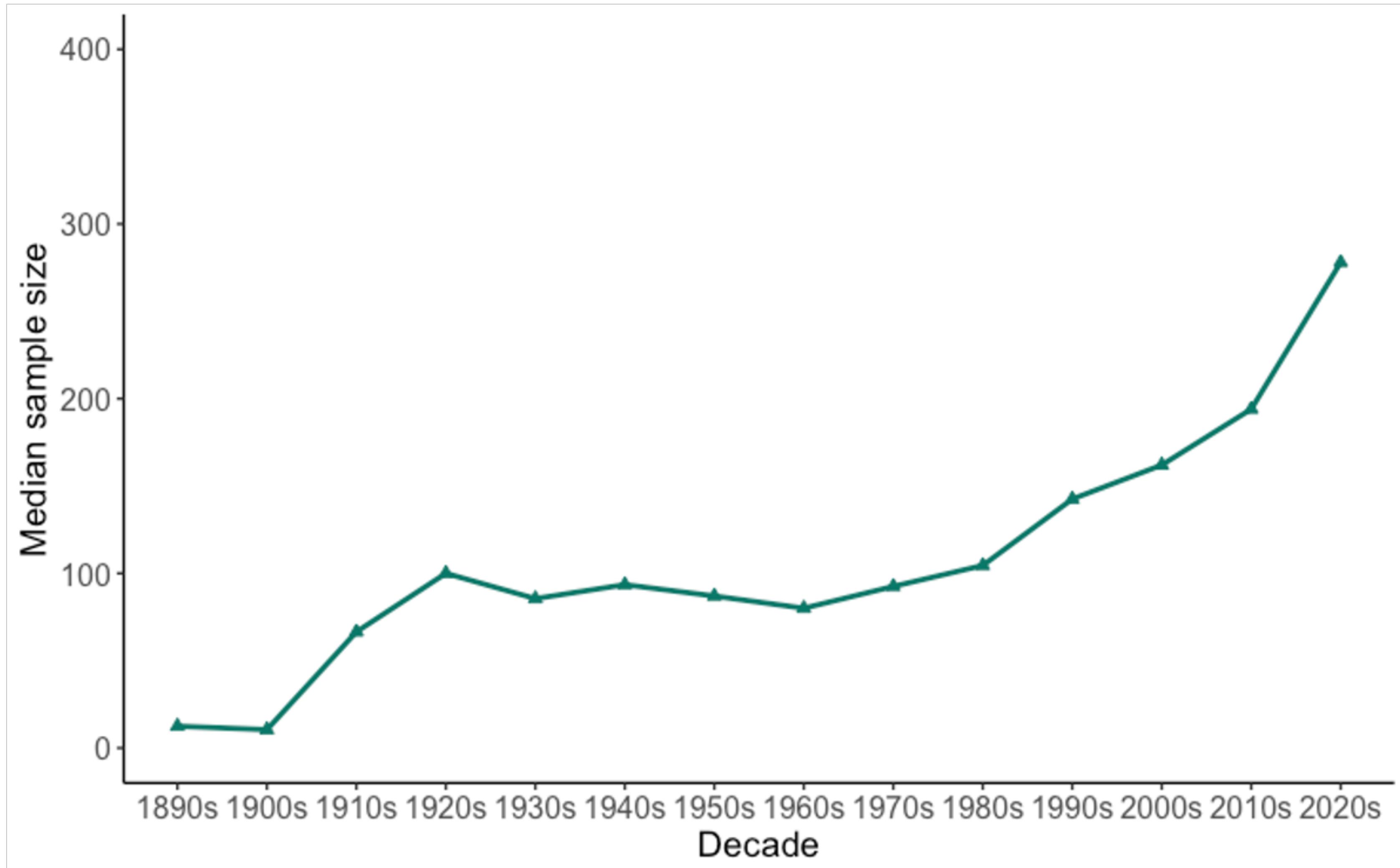
Participant Characteristics

	Characteristic	Value
1	Sample Size	1205
2	Gender (Female)	50%
3	Age (Mean)	47.7
4	Age (SD)	16.6
5	Age Range	18-87

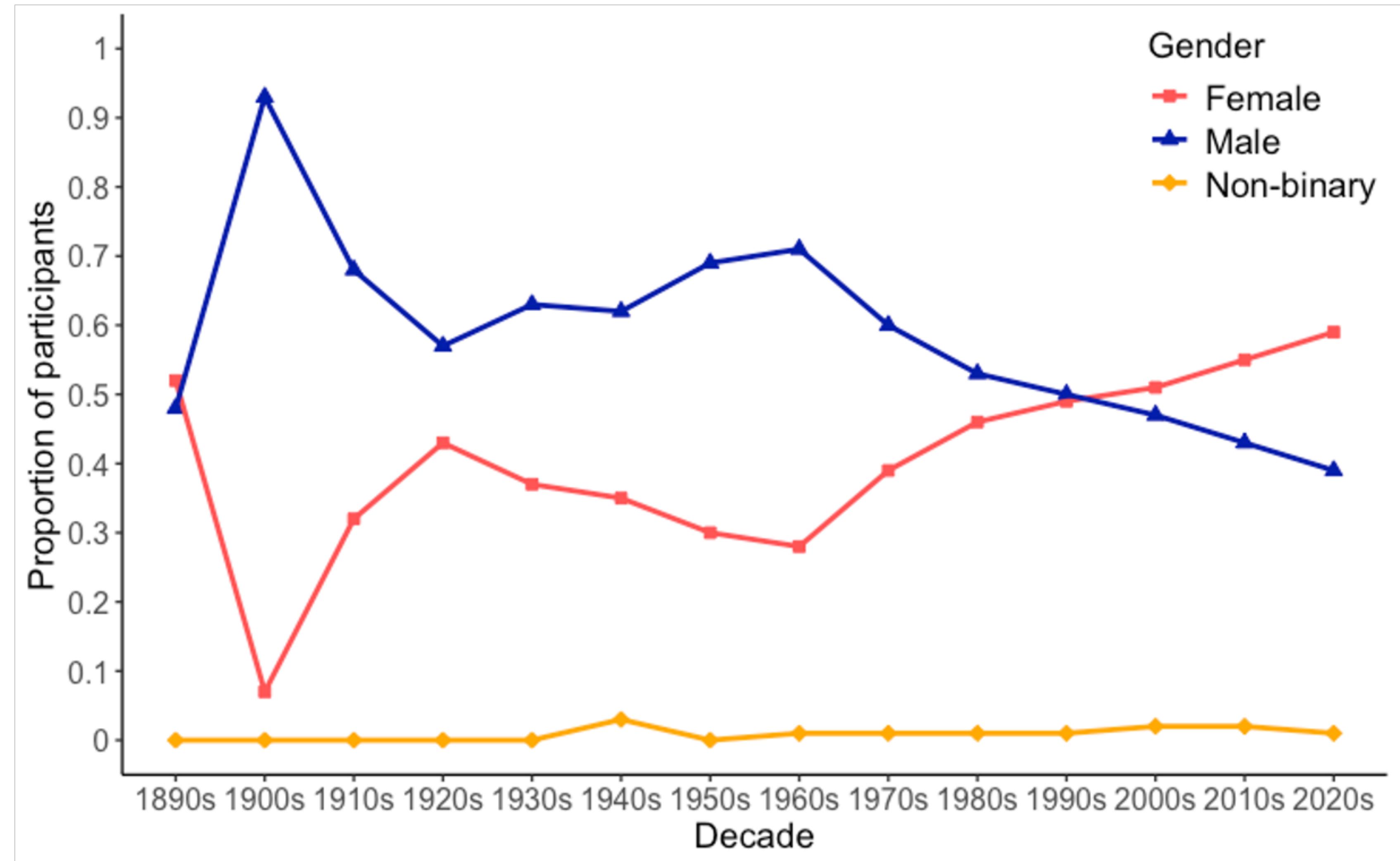
The table of participant characteristics, including demographic information for the studies included in the document, has been created and displayed for you. If you need any additional information or further analysis, please let me know! [\[?\]](#)



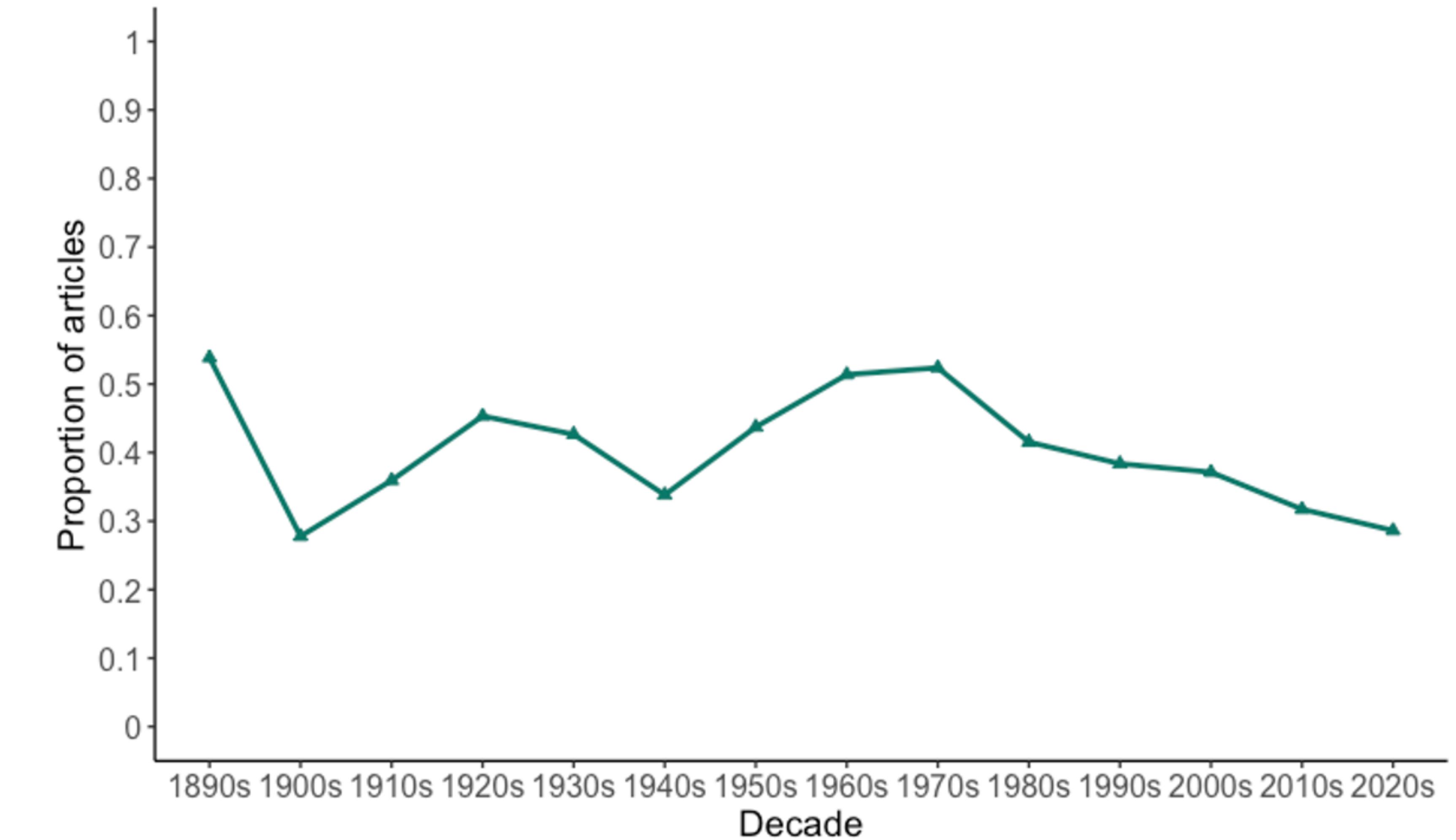
Sample sizes increase drastically



Samples are increasingly female

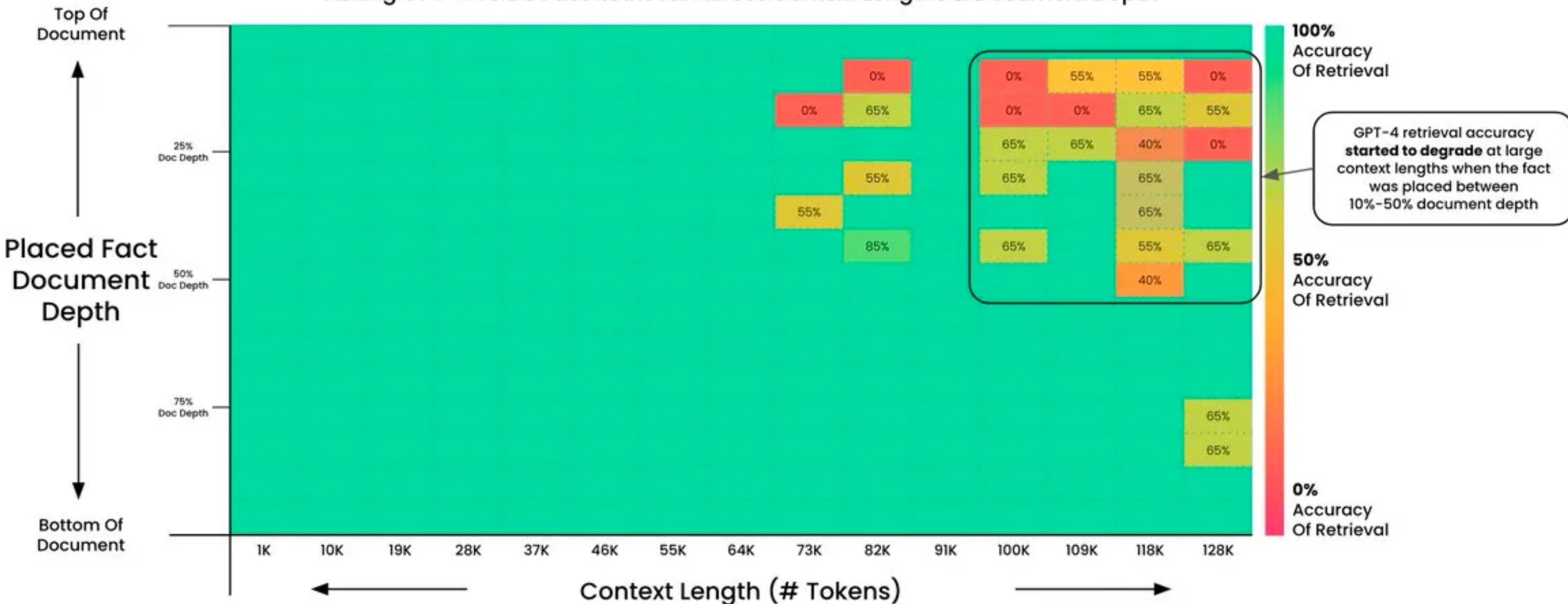


Still many samples are drawn from student populations



Pressure Testing GPT-4 128K via "Needle In A HayStack"

Asking GPT-4 To Do Fact Retrieval Across Context Lengths & Document Depth



Goal: Test GPT-4 Ability To Retrieve Information From Large Context Windows

A fact was placed within a document. GPT-4 (1106-preview) was then asked to retrieve it. The output was evaluated for accuracy.

This test was run at 15 different document depths (top > bottom) and 15 different context lengths (1K > 128K tokens).

2x tests were run for larger contexts for a larger sample size.

Exercise

Talk to documents

Against Justaism: A call for more measured discussions on LLM cognition

Zak Hussain^{1,2*}, Rui Mata¹ and Dirk U. Wulff^{2,1}

¹*University of Basel.
²Max Planck Institute for Human Development.

*Corresponding author(s). E-mail(s): z.hussain@unibas.ch;
Contributing authors: rui.mata@unibas.ch; wulff@mpib-berlin.mpg.de;

Abstract

Are large language models (LLMs) "just next-token predictors" devoid of cognitive capacities such as "thought" or "understanding"? We argue from a cognitive science perspective that it is far from self-evident that such perspectives hold despite their prominence in current discussions surrounding LLMs. We make our case by critically evaluating different flavors of *Justaic* reasoning and end with a call for more measured discussions on LLM cognition.

Keywords: large language models, cognition

