Large Language Models as Computational Instruments for Social Science Research

Max Pellert (https://mpellert.at)

[Slides as PDF]

Classical traditional surveys

Beloved standard tool of the social sciences

Often considered the gold standard, "ground truth" (especially when working with large representative samples of a population)

But, classical survey methodologies increasingly suffer from problems

First line of the 2024 Book "Polling at a Crossroads: Rethinking Modern Survey Research": *Survey research is in a state of crisis*

Last example: US presidential elections 2024

Biggest issue is non-response

Alternatives?

Synthetic Surveys

Britain's mood, measured weekly

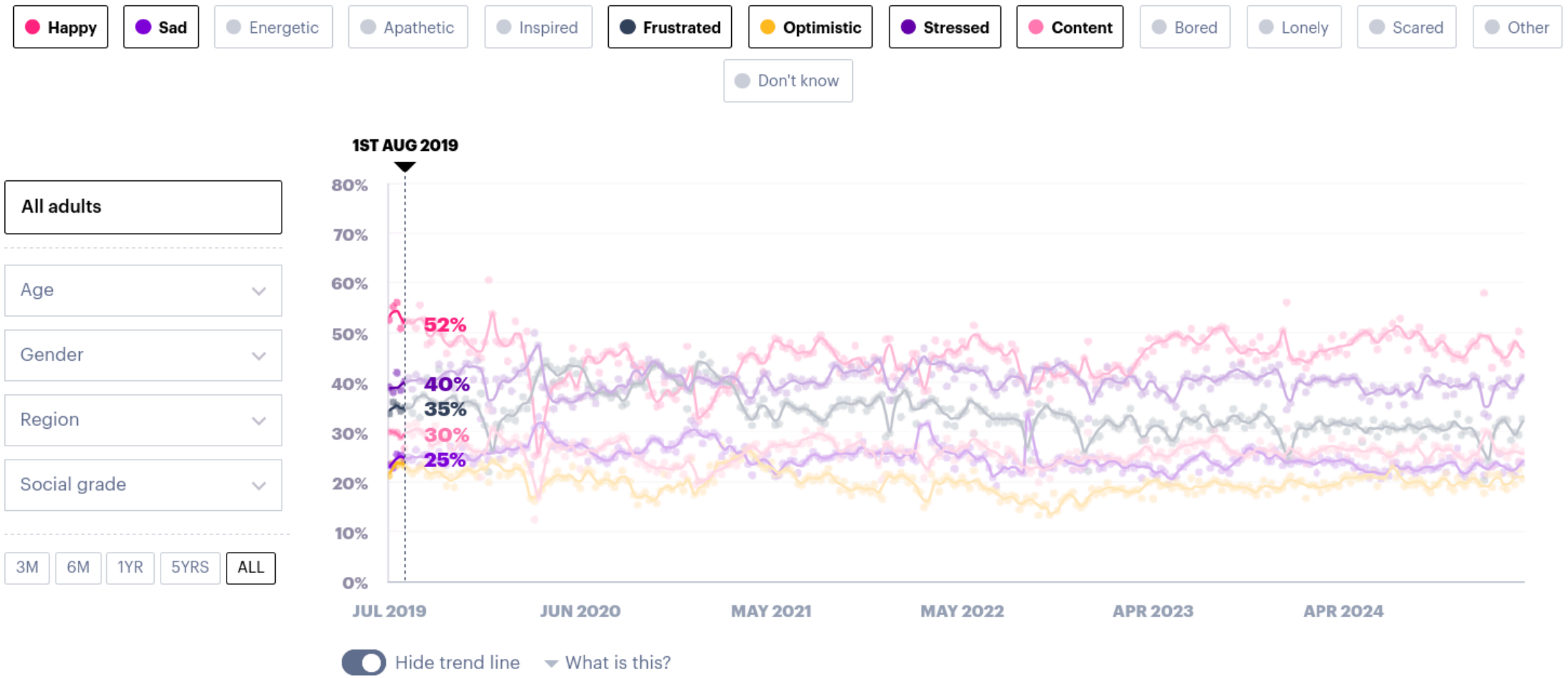
One example of an easily accessible, representative survey (UK) in the affective domain

https://yougov.co.uk/topics/politics/trackers/britains-mood-measured-weekly

Ahnert, G., Pellert, M., Garcia, D., & Strohmaier, M. (2024). Britain's Mood, Entailed Weekly: In Silico Longitudinal Surveys with Fine-Tuned Large Language Models. Companion Proceedings of the 16th ACM Web Science Conference, 47–50. https://doi.org/10.1145/3630744.3659829

Ahnert, G., Pellert, M., Garcia, D., & Strohmaier, M. (2025). Extracting Affect Aggregates from Longitudinal Social Media Data with Temporal Adapters for Large Language Models. Proceedings of the International AAAI Conference on Web and Social Media, 19, 15–36. https://doi.org/10.1609/icwsm.v19i1.35801

Britain's mood, measured weekly



FULL QUESTION

Broadly speaking, which of the following best describe your mood and/or how you have felt in the past week Please select all that apply

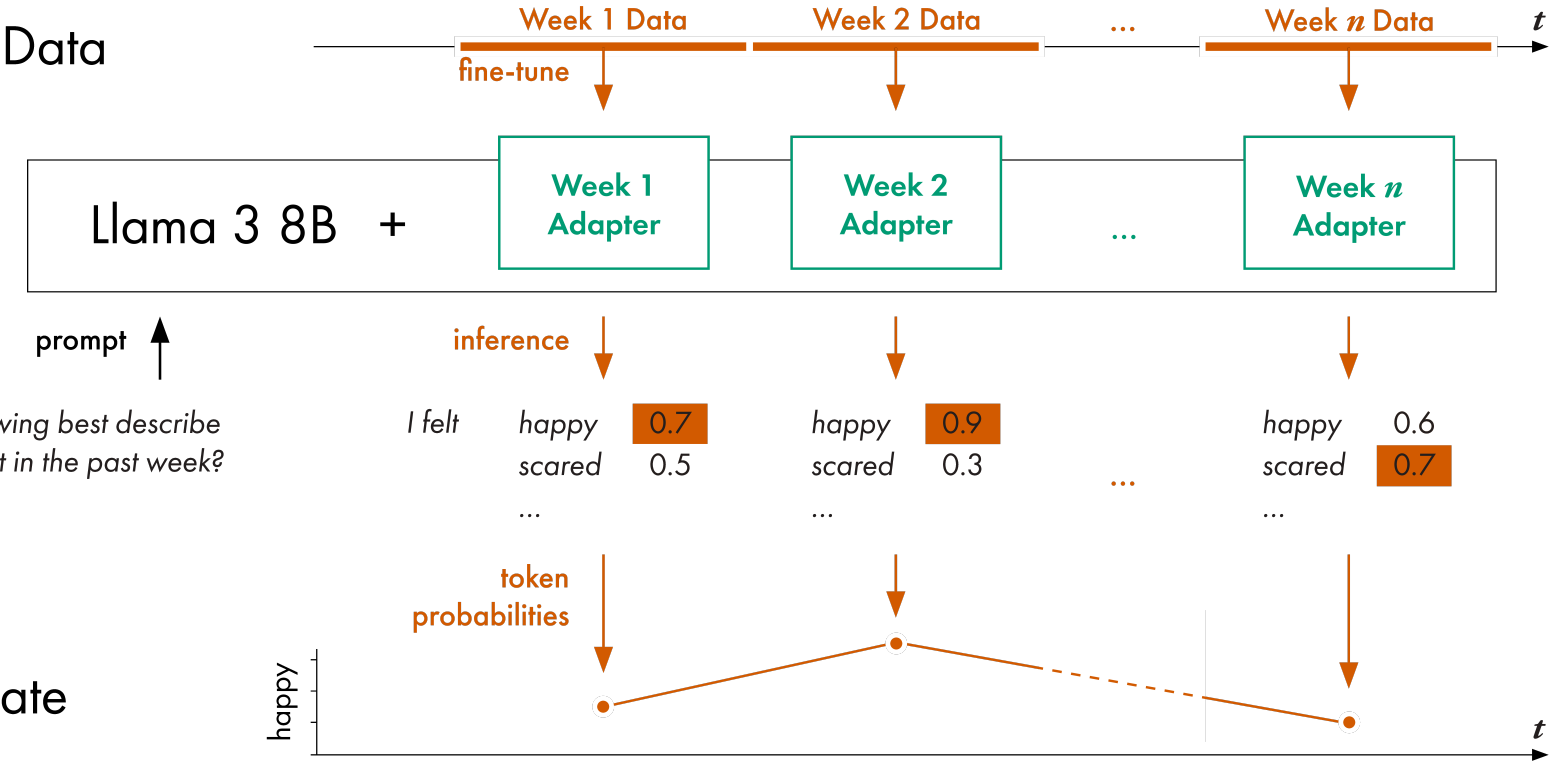
Weekly Social Media Data

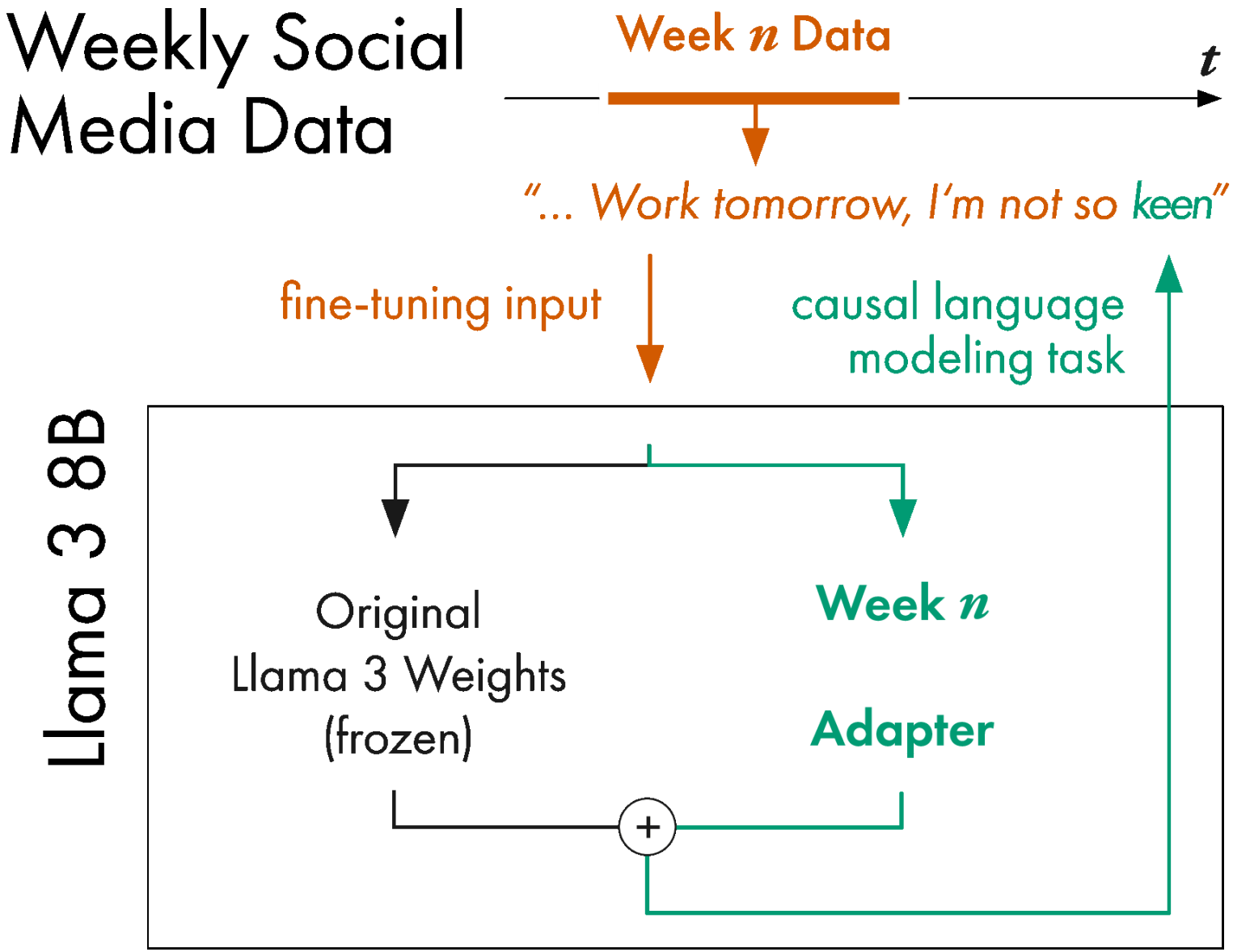
Temporal Adapters

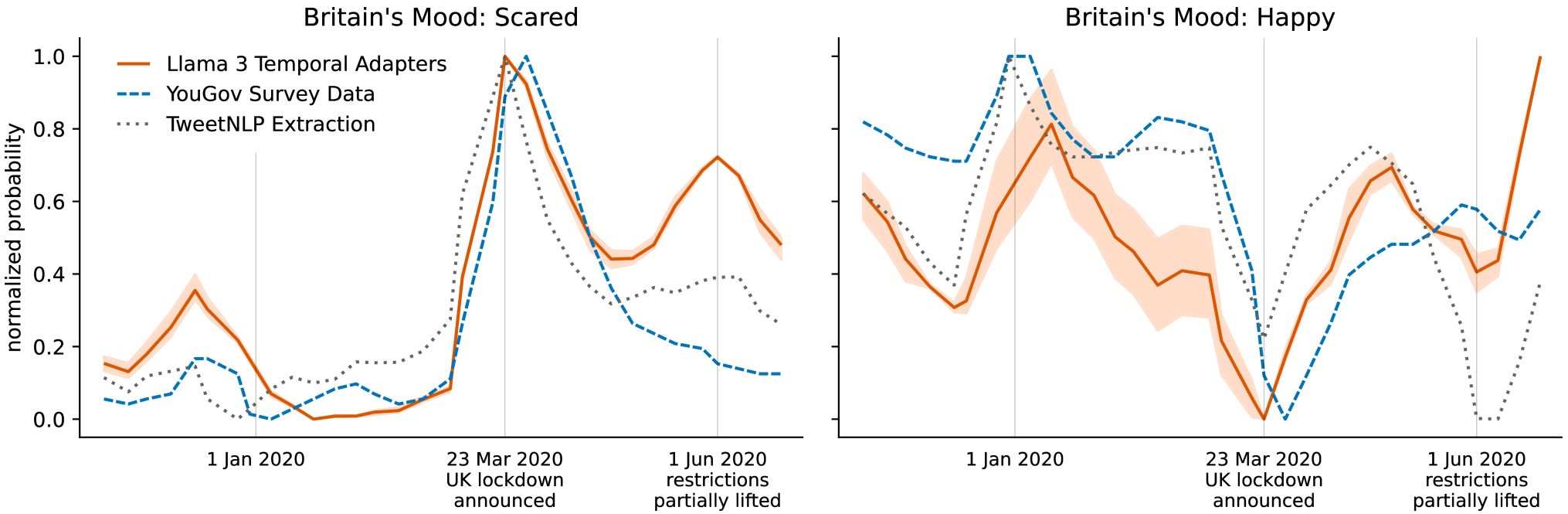
Survey Question

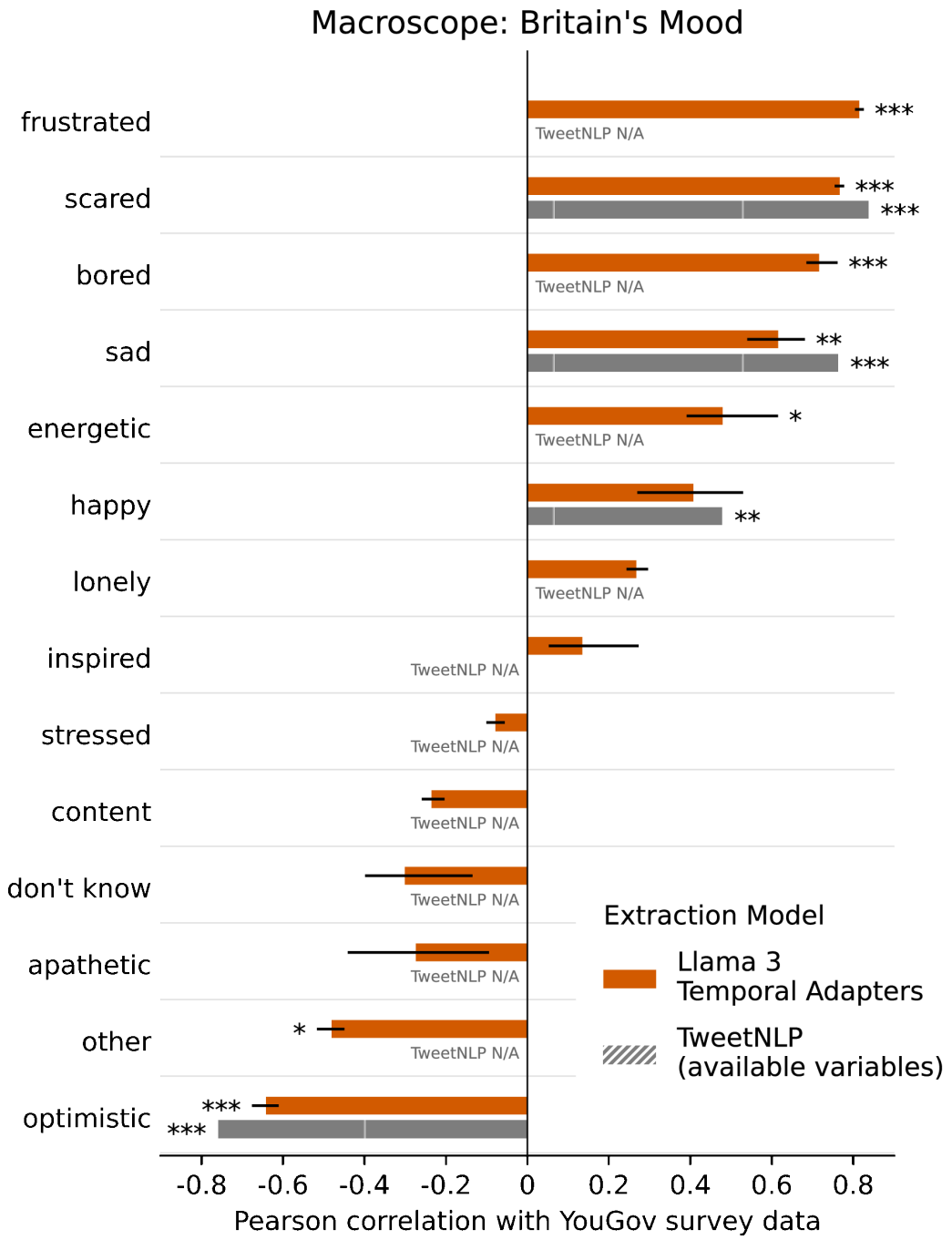
Broadly speaking, which of the following best describe your mood and/or how you have felt in the past week?

Weekly Affect Aggregate




Weekly Social Media Data
Week n Data
t
"... Work tomorrow, I'm not so keen"
fine-tuning input
causal language modeling task
Llama 3 8B
Original Llama 3 Weights (frozen)
Week n
Adapter
+


Britain's Mood: Scared
Britain's Mood: Happy
Llama 3 Temporal Adapters
YouGov Survey Data
TweetNLP Extraction
normalized probability
1.0
0.8
0.6
0.4
0.2
0.0
1 Jan 2020
23 Mar 2020
UK lockdown
announced
1 Jun 2020
restrictions
partially lifted


Macroscope: Britain's Mood
frustrated
scared
bored
sad
energetic
happy
lonely
inspired
stressed
content
don't know
apathetic
other
optimistic
***
TweetNLP N/A
***
***
***
TweetNLP N/A
**
***
*
TweetNLP N/A
**
TweetNLP N/A
TweetNLP N/A
TweetNLP N/A
TweetNLP N/A
TweetNLP N/A
*
TweetNLP N/A
***
***
Extraction Model
Llama 3
Temporal Adapters
TweetNLP
(available variables)
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
Pearson correlation with YouGov survey data

Results

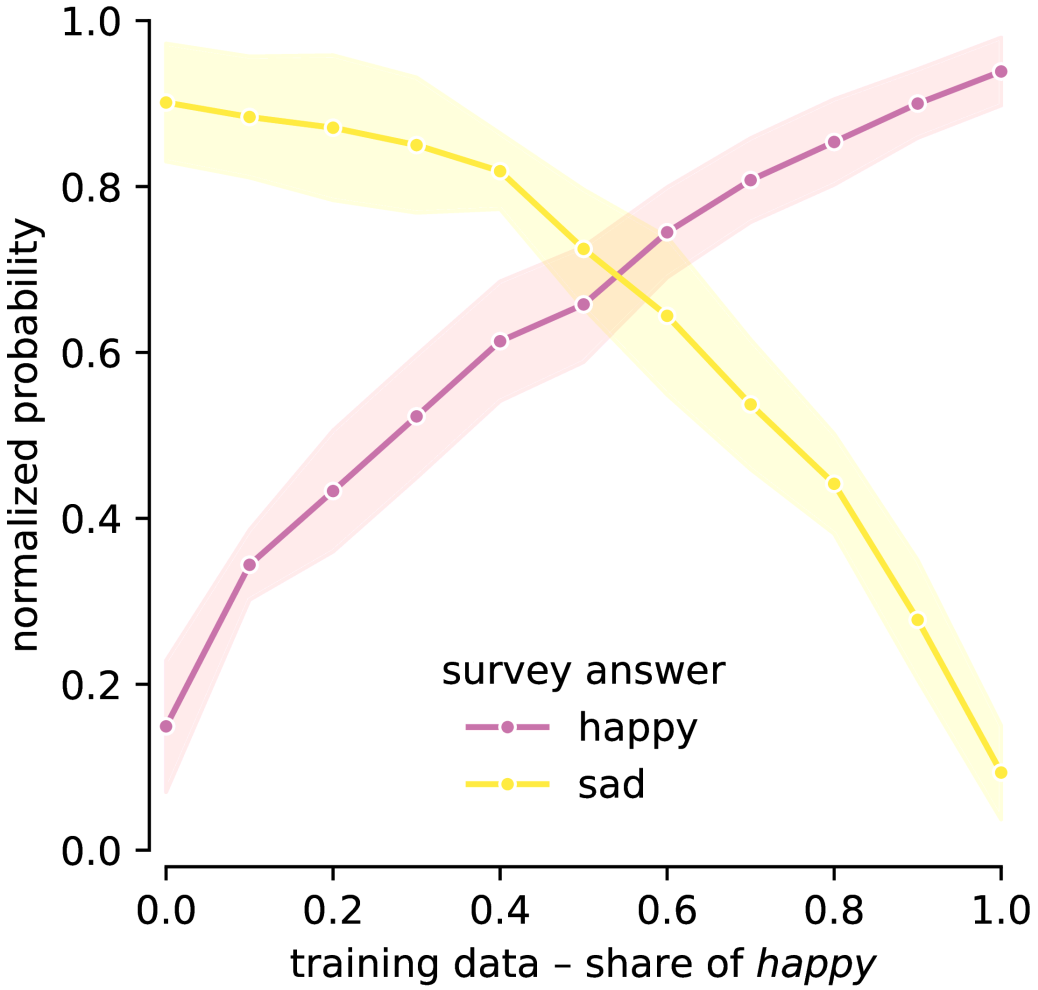
We can recreate dynamics with that approach of longitudinal adaptors

Not equally well for all constructs

Remember, our approach is just self-supervised next token prediction (no labels present as for example with the supervised text classification method of TweetNLP)

Our approach is very flexible, we can in principle ask any question and get survey-like responses for each week

Why does that work?


1.0
0.8
0.6
0.4
0.2
0.0
normalized probability
survey answer
happy
sad
0.0
0.2
0.4
0.6
0.8
1.0
training data – share of *happy*

Wrap-up

I don't think we should be replacing survey research

Also with complementary synthetic methods we will need classical approaches for example to learn about the sampling frame

But we should be making use of the text that people are producing (and potentially other modalities too)

It's first steps for now and we strongly have to validate what we are doing

Huge potential: Low costs, scalability, unobtrusive observation, high temporal resolution, …

Bridging the gap between "qualitative" data and quantitative insights

Next: LLMs for Digital Twinning

LLMs increasingly deployed as autonomous agents

Research gap: alignment with actual human decision-making

Why Game Theory?

Analytical solutions (Nash equilibria) as benchmarks

Rich empirical data from human experiments

Simple, well-defined tasks

Real-world relevance

Goal: Replication of human experimental data with LLMs, systematically validated → novel predictions

Poncela-Casasnovas, J., Gutiérrez-Roig, M., Gracia-Lázaro, C., Vicens, J., Gómez-Gardeñes, J., Perelló, J., Moreno, Y., Duch, J., & Sánchez, A. (2016). Humans display a reduced set of consistent behavioral phenotypes in dyadic games. Science Advances, 2(8), e1600451. https://doi.org/10.1126/sciadv.1600451

Methods

Models: Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3, Qwen2.5-7B-Instruct

Original Experiment: 500+ humans, 121 games (Human behavioral phenotypes across games: **All deviate from Nash equilibrium**)

Payoff Structure:

	C	D
C	(10,10)	(S,T)
D	(T,S)	(5,5)

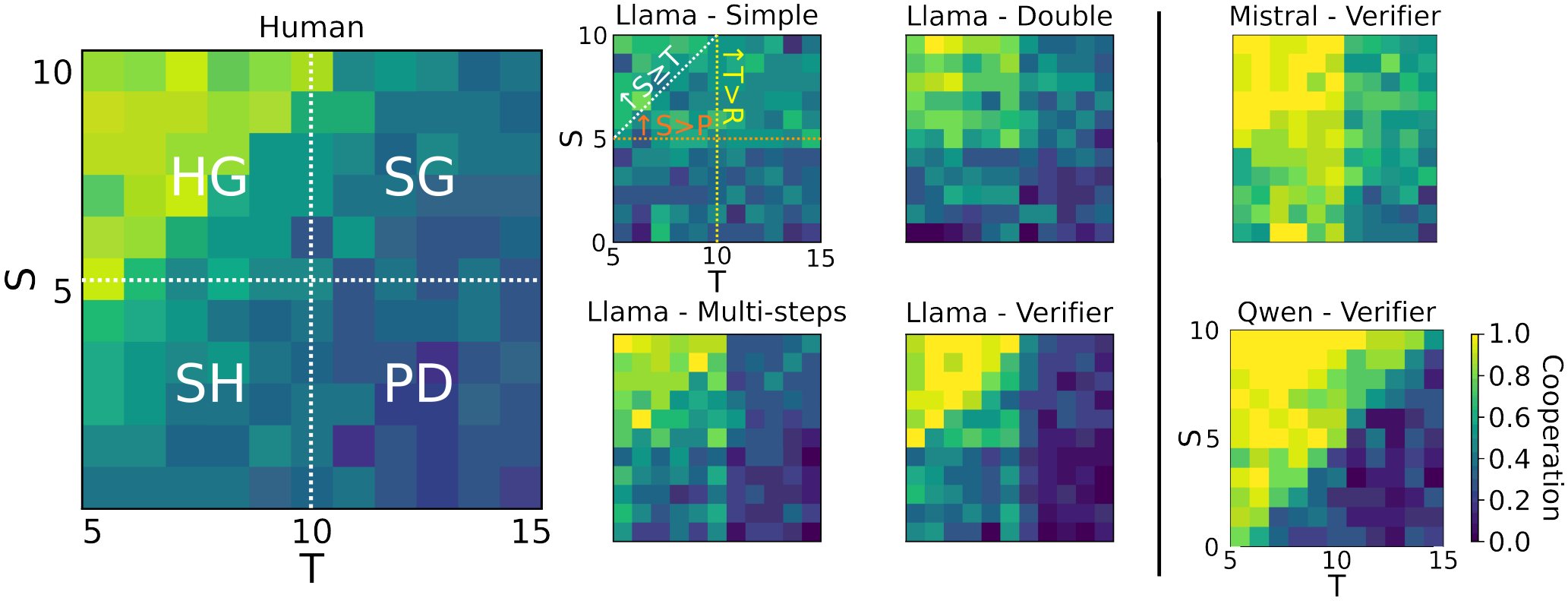
$S \in [0,10]$, $T \in [5,15] \rightarrow$ Extended through our simulations to [0,20]

Progressive Answer Extraction

1. **Simple:** Direct answer → (mostly) random patterns
2. **Double:** Long answer + extraction → some structure
3. **Multi-step:** Guided reasoning → clear patterns
4. **Logical Verifier:** + validation → high algorithmic fidelity

"Thinking step-by-step" improves coherence

Logical verifier acts as an "LLM attention check", specifically checking on consistency in the Harmony Game region


Human
HG
SG
SH
PD
S
T
Llama - Simple
S=T
T>R
S>P
Llama - Double
Mistral - Verifier
Llama - Multi-steps
Llama - Verifier
Qwen - Verifier
Cooperation

Quantitative Model Comparison

	Human		Nash	
	MSD	r	MSD	r
Llama	**0.031**	**0.89**	0.089	0.77
Mistral	0.091	0.70	0.182	0.60
Qwen	0.065	0.79	**0.036**	**0.93**
Nash	0.096	0.78	-	-

Llama replicates humans (better than Nash); Qwen follows Nash; Mistral intermediate

Observations

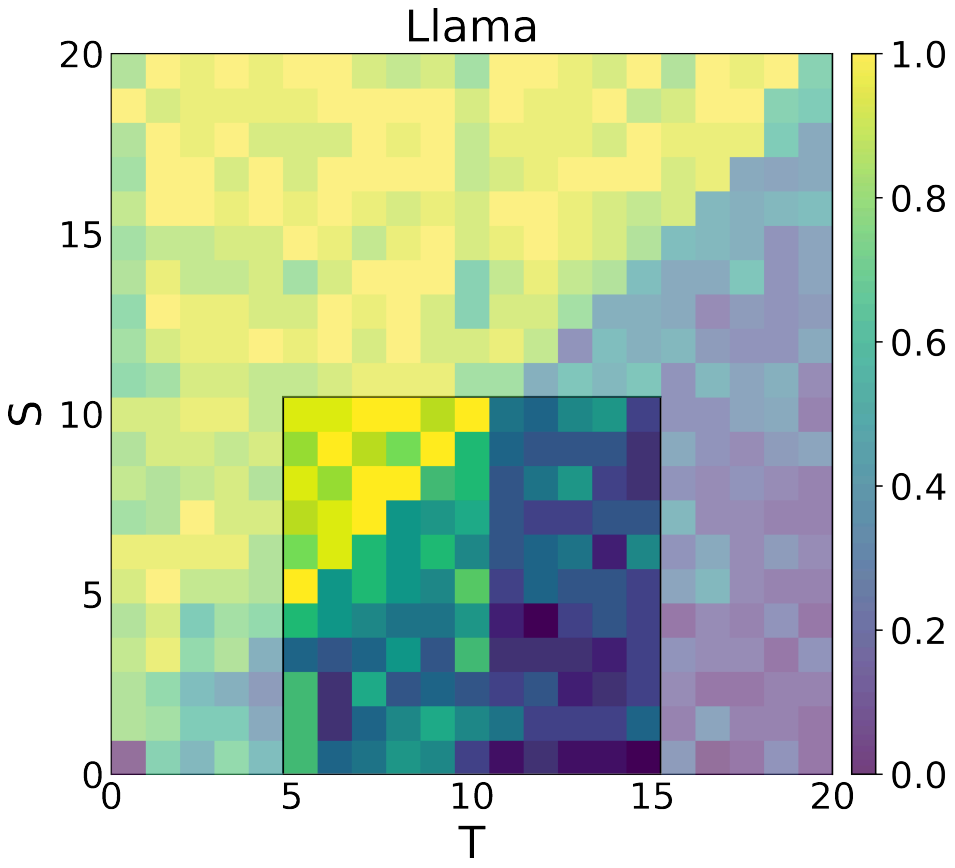
Human vs. Llama similarities:

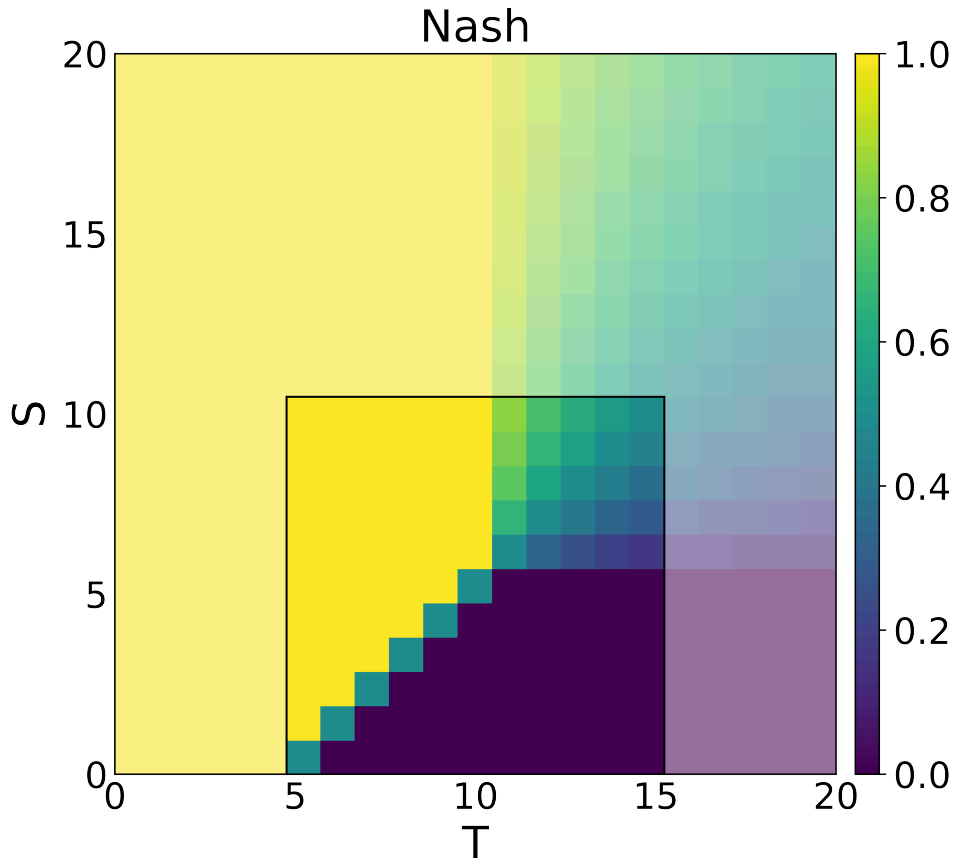
- High cooperation when $S \geq T$
- Low cooperation when $T > R$
- Binary-like patterns

Llama (and human) vs Nash:

- No mixed equilibria
- Discrete choices
- Emulating (human) psychological heuristics?

Average cooperation: Llama 40.2%, Human 48.0% vs. Nash prediction of 50%


Llama
S
T
0
5
10
15
20
0.0
0.2
0.4
0.6
0.8
1.0


Nash
S
T
0
5
10
15
20
0.0
0.2
0.4
0.6
0.8
1.0

Novel Game Predictions

Extended 121 → 441 games

Llama patterns beyond human-tested space:

- $S \geq T$ diagonal holds
- $T > R$ reduces cooperation
- Instability near (0,0)

Pre-registered experiment for future validation[1]

1. https://aspredicted.org/fe6z2k.pdf

Key Contributions

- Population-level replication without personas
- Open-source models (reproducible)
- Logical verification as quality control
- Outperforms Nash at predicting humans (Training creates **behavioral imitators**)
- Generates testable hypotheses

Limitations

- Edge case instability
- Potential memorization concerns
- Black-box mechanisms
- Requires human validation

Conclusions and Implications

With the right protocol, we can use LLMs to replicate human patterns and to capture deviations from rationality

Complementary tool for the social and social and behavioral sciences

Rapid experimental space exploration

Generate hypotheses → validate with humans

AI-assisted scientific discovery

Pre-Print: Palatsi, A. C., Martin-Gutierrez, S., Cardenal, A. S., & Pellert, M. (2025). Large language models replicate and predict human cooperation across experiments in game theory (arXiv:2511.04500). arXiv. https://doi.org/10.48550/arXiv.2511.04500

Code: github.com/acerapal/Replicating-Human-Game-Theory-Experiments-with-LLMs

Next: AI Psychometrics

The Promise and the Problem

Psychometrics depends on **human rater judgments** to measure latent constructs (values, personality, attitudes)

Questionnaire development is **iterative and costly**: large candidate item sets, extensive piloting, respondent fatigue, low motivation

Growing interest in using **NLP and embeddings** to complement psychometric workflows

But: Existing embedding approaches face a critical limitation: they cannot produce **negative correlations** between items

Test Case: Schwartz's Value Theory

Values are **trans-situational goals** that guide behavior

The Revised Portrait Value Questionnaire (PVQ-RR): 57 items, **19 fine-grained value dimensions**

Values form a **circumplex structure**: motivationally compatible values correlate positively and are close, opposing values correlate negatively and are far apart

Cross-culturally validated with human data from **49 countries** (Schwartz & Cieciuch, 2022)

Why is this hard? The circular structure depends critically on **negative correlations** between opposing value dimensions, which is exactly what embeddings struggle with

Prior Work and the Missing Negatives

Cutler & Condon (2023): Template-based approach with BERT, limited to personality descriptors

Wulff & Mata (2025): Domain-specific finetuning on 200,000 item pairs, dropped sign of correlation entirely

Hommel & Arslan (2025): Re-annotated NLI corpus to handle negation, then finetuned

The root cause: High-dimensional embeddings encode many shared abstract linguistic features that **artificially inflate similarity** between all items, even semantically opposing ones

SQuID: Our Methods Contribution

Survey and **Qu**estionnaire **I**tem Embeddings **D**ifferentials

For each item embedding $\mathbf{y}_i$, subtract the mean embedding over all items:

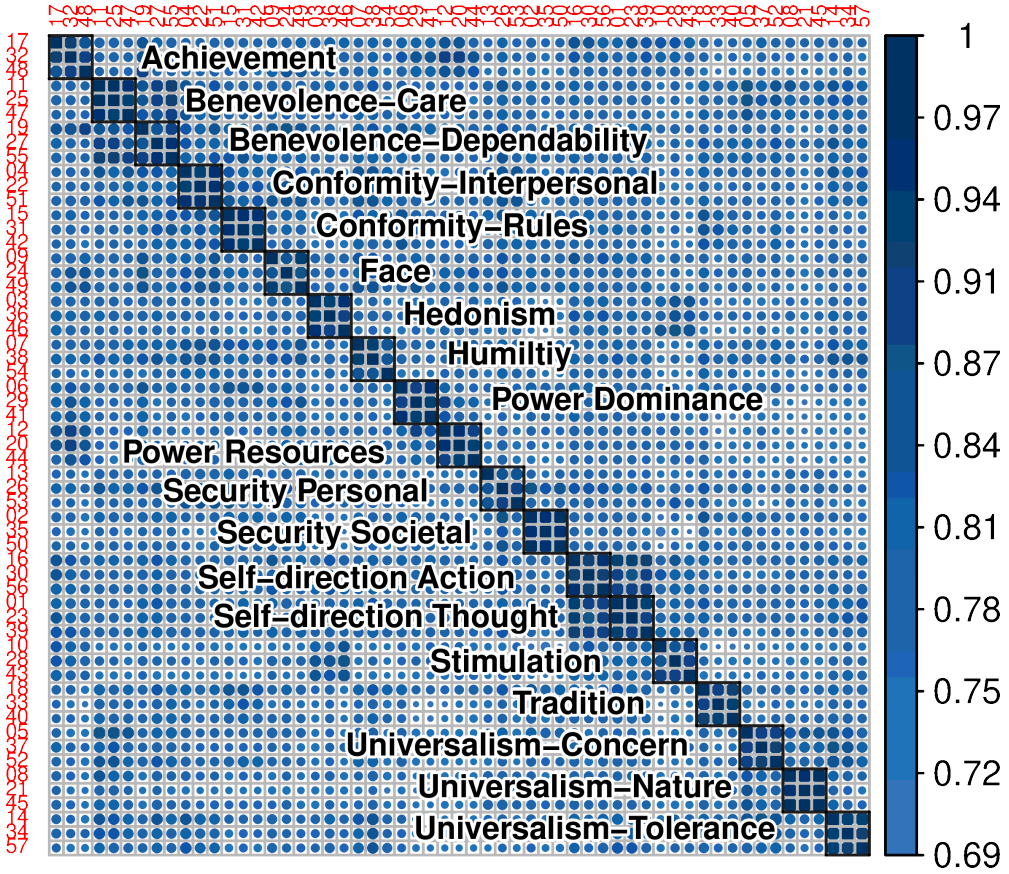
$$\mathbf{y}_i - \bar{\mathbf{y}}, \quad \text{where } \bar{\mathbf{y}} = \frac{1}{57} \sum_{i=1}^{57} \mathbf{y}_i$$

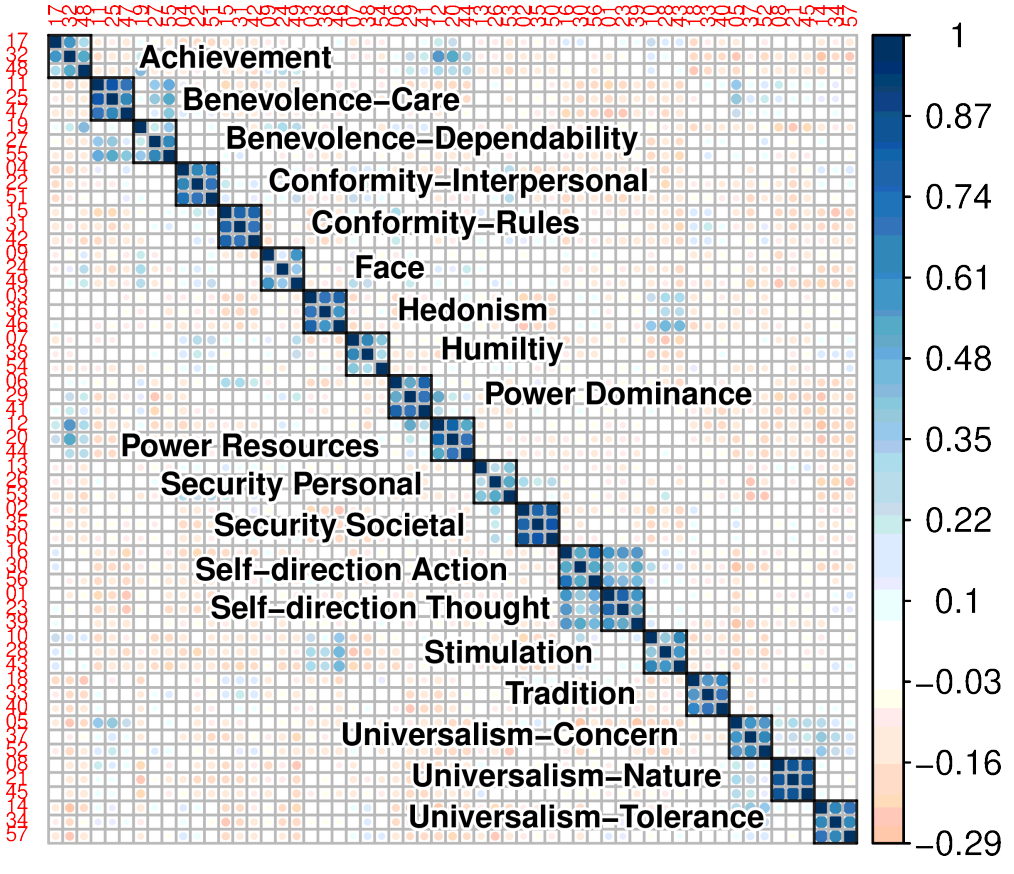
Removes shared linguistic features that inflate similarity

Inspired by Word2Vec arithmetic (Mikolov et al., 2013) and community embeddings (Waller & Anderson, 2021)

Works with **any model**, **any embedding dimension**, requires **no finetuning**, no re-annotation, applied purely **post-hoc**

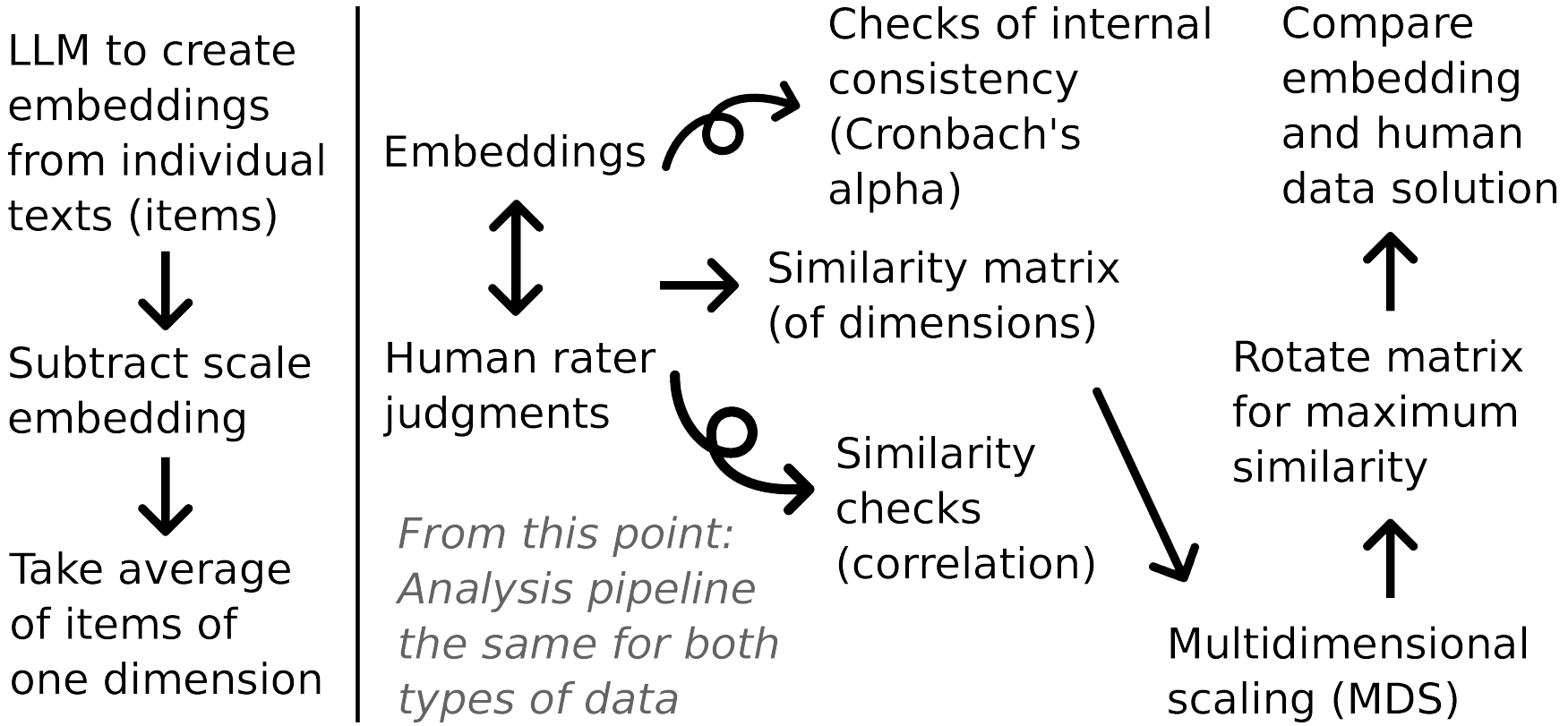
The Effect of SQuID





By subtracting the questionnaire embedding, negative correlations clearly appear without any domain-specific adaptation

Workflow Overview



From the similarity matrix onward, the analysis pipeline is **identical** for embeddings and human rater judgment data

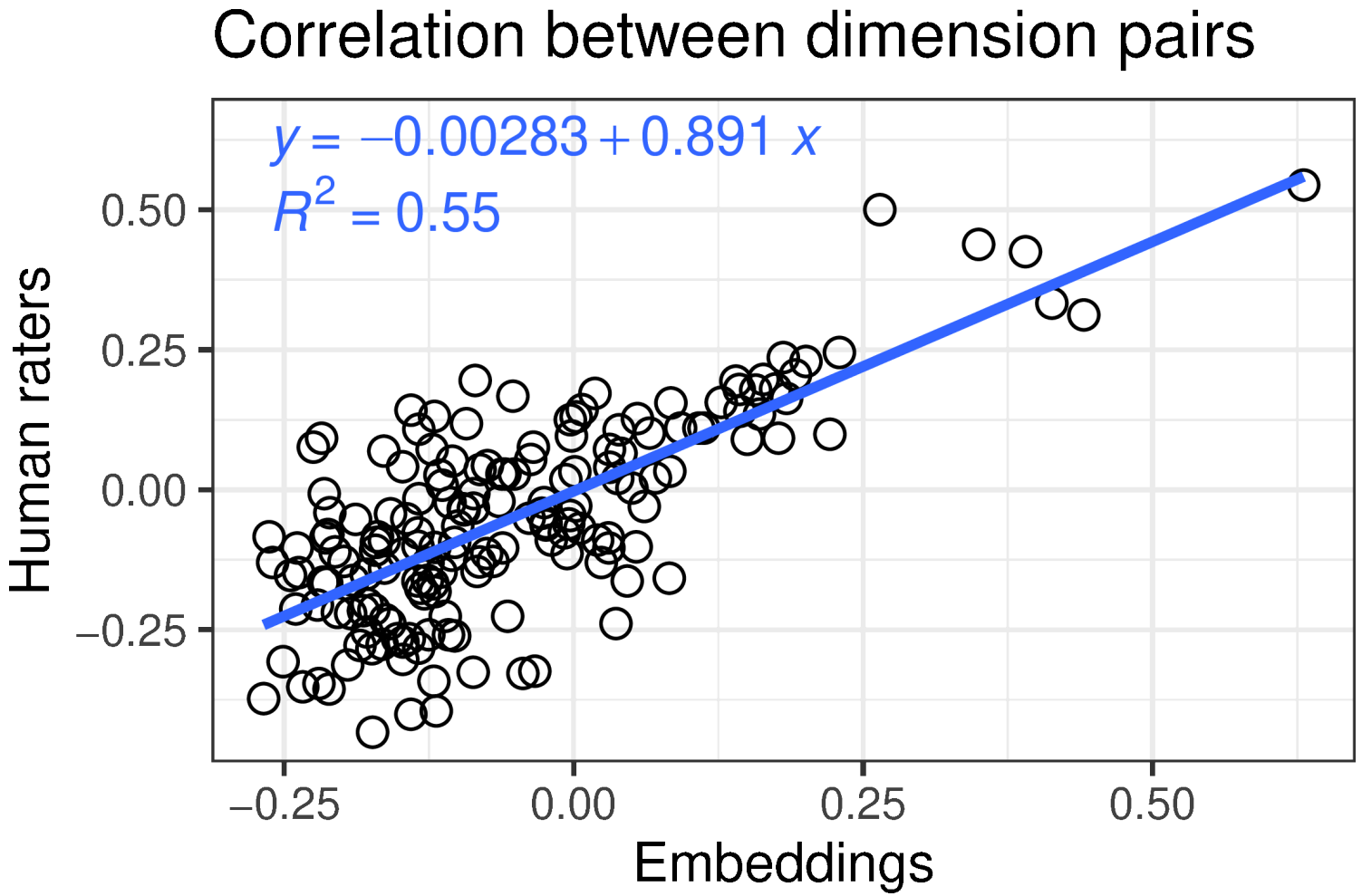
Internal Consistency: Cronbach's Alpha

	Average alpha
Linq-Embed-Mistral	**0.77**
Human data (49 countries)	0.70
gemini	0.64
jina / kalm	0.67
mpnet-personality	0.57
Random baseline	-0.05

Linq-Embed-Mistral **beats human data** on average Cronbach's alpha and wins **14 out of 19** individual dimensions

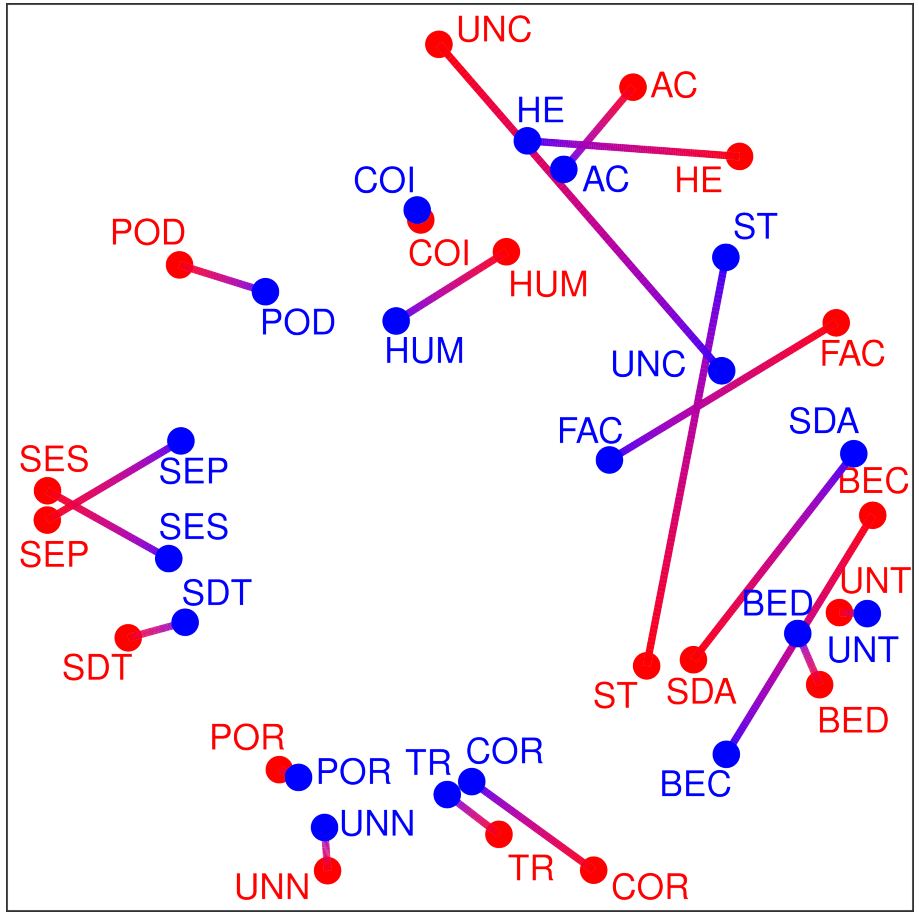
MTEB benchmark ranking serves as a good guide for model selection in this task

Dimension-Dimension Correlations



171 dimension pairs: r = 0.74 [0.66, 0.80], simple linear model explains **55% of variance**, no bias

MDS: Circular Value Structure Emerges



Facets of the same construct cluster together, opposing values are placed on opposite sides

Generalizability Beyond Value Theory

SQuID tested on three personality questionnaires: **IPIP**, **BFI-2**, **HEXACO**

Questionnaire	Avg. relative gain in range
IPIP	261.5%
BFI-2	207.2%
HEXACO	147.9%

Individual model gains range from 21% (mpnet, already domain-finetuned) to 724% (kalm)

Even domain-finetuned models benefit from SQuID. Consistent improvements suggest **broad applicability**.

Discussion and Future Directions

What we showed: SQuID enables embeddings to recover psychometric structures on par with human data, with no finetuning and broad generalizability

Exciting future direction: Systematic assessment of **cultural biases** in LLMs by comparing embedding MDS configurations per country with human data from 49 countries

Open challenges: Reverse-keyed items, hierarchical and network-based psychometric models, multilingual evaluation, clinical inventories

Our vision: Not merely another step in the validation pipeline, but a new way to investigate the full spectrum of human behavior and experience across all natural languages

The end

Max Pellert

max.pellert@bsc.es

Barcelona Supercomputing Center

Code and data: github.com/maxpel/embeddings_values