Large Language Models as Social Scientific Objects and Instruments

Measurement, Validation and Prediction

Max Pellert

[Slides as PDF]

Social data science – and where it leads

My work is rooted in large-scale empirical social data: signed networks from 400M+ interactions (DerStandard corpus), longitudinal social media data, sentiment analysis and emotion detection

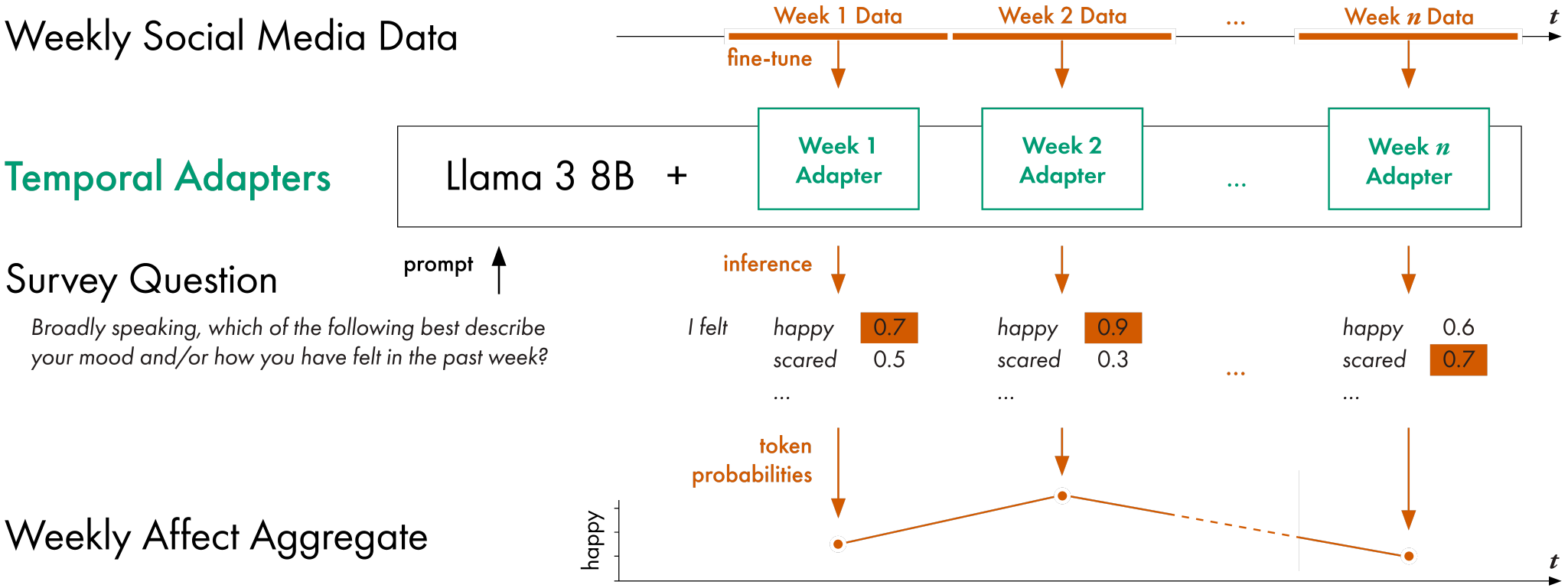
I held the **Professor for Social and Behavioral Data Science** chair at the University of Konstanz, and I have researched and taught data science at Mannheim, Konstanz, and in industry contexts

Increasingly, the most consequential social data is not just generated *by* humans rather is it generated *by AI systems interacting with humans*

That shift is where my research program lives: can we use and study AI systems and maintain the same empirical rigor we apply to human subjects?

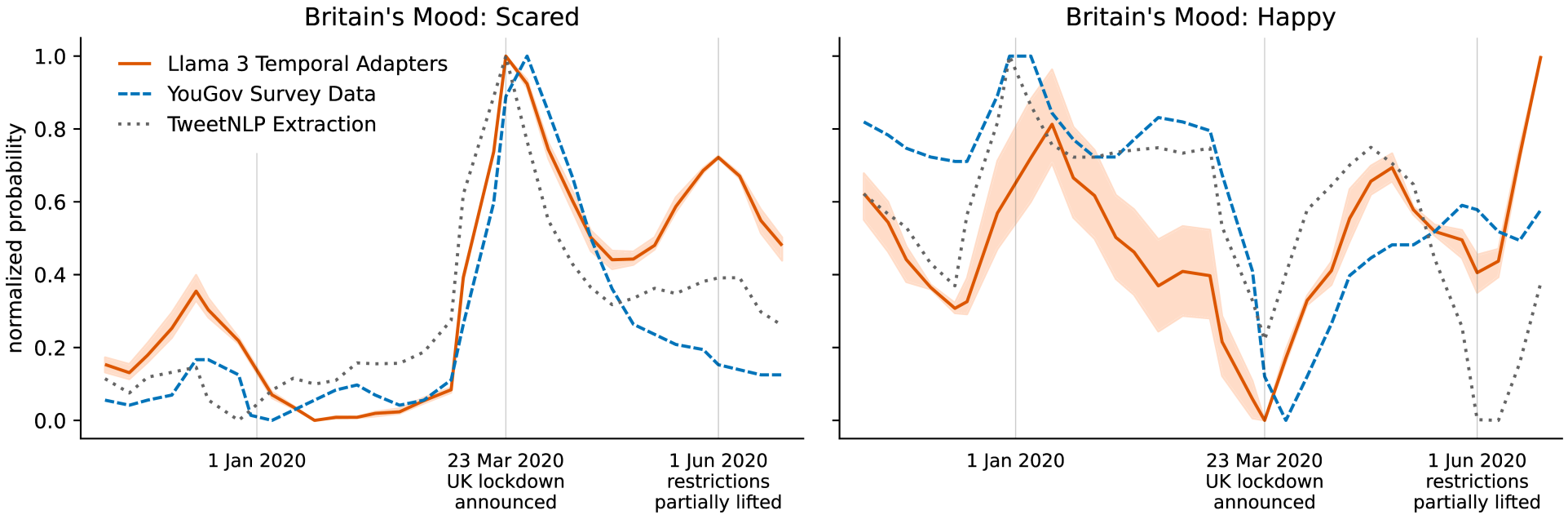
One concrete example from our work first: using LLMs as instruments to recover what a society is feeling, at scale and over time

Synthetic surveys: LLMs as social instruments



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

Validated against representative survey data



Strong positive correlations with weekly YouGov survey data across multiple collective emotions during COVID-19 (robust across training seeds and prompt formulations)

Key property: flexibility, we can ask *any* question, including ones for which no traditional classifier building on annotated training data exists

Can LLMs replicate human social behavior?

121 two-player games spanning the full space of social dilemmas: Prisoner's Dilemma, Stag Hunt, Harmony, Chicken, and all transitions between them

Empirical baseline: 500+ human participants from a landmark behavioral science study (Poncela-Casasnovas et al., Science Advances 2016)

Three open-source models: Llama-3.1-8B, Mistral-7B, Qwen2.5-7B

What we found: models have stable, reproducible behavioral profiles that differ systematically – one tracks human behavior closely, one follows Nash equilibrium, one falls in between

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

The experimental space

Payoff Structure:

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

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

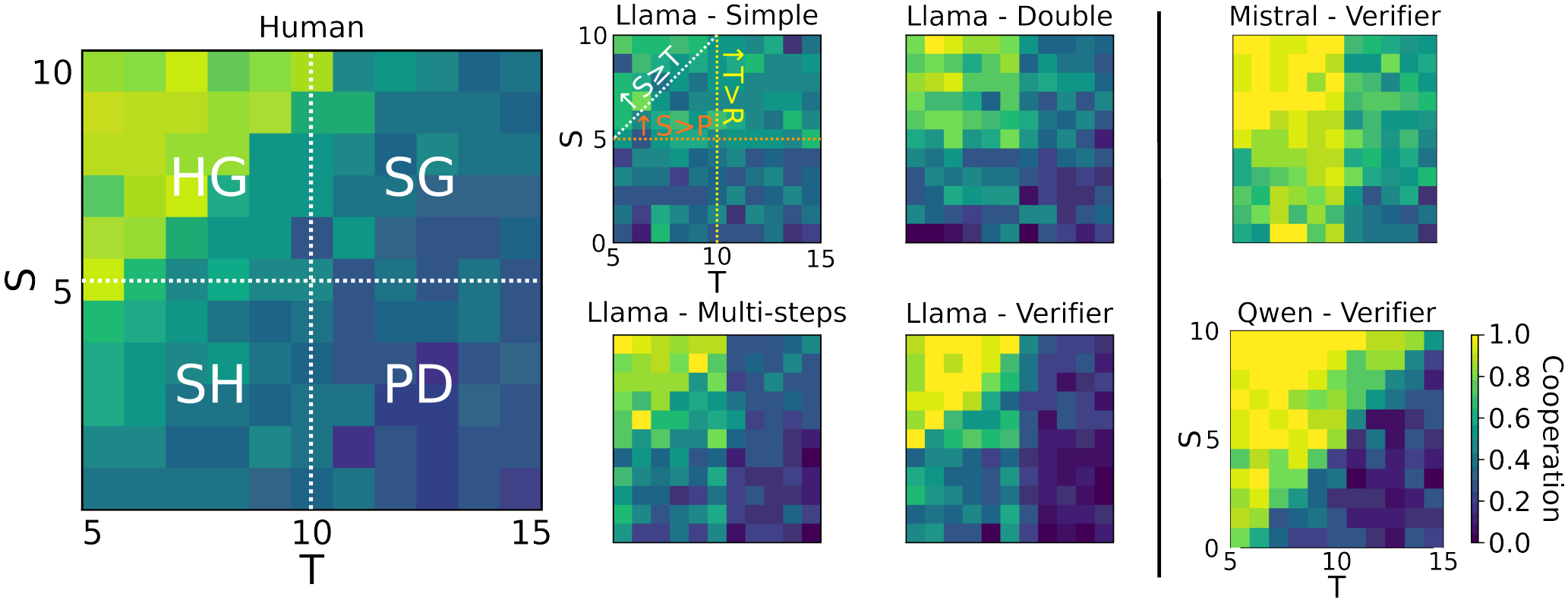
Scale: 441 game conditions × 3 models × 20 iterations × multiple prompting strategies creates a large behavioral dataset in the classical experimental social science sense, systematically compared against human ground truth

The experimental space

The game theoretic framework we are using allows us also to compare expected cooperation rates building on assumptions of rationality (Nash equilibria), derived both analytically and from simulations in evolutionary game theory

Progressive Answer Extraction for LLMs needed: four increasingly complex layered prompting strategies to elicit structured responses – from direct answers to multi-step reasoning with logical verification

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


Human
HG
SG
SH
PD
S
T
10
5
15
Llama - Simple
↑S=T
↑T=R
↑S>P
Llama - Double
Mistral - Verifier
Llama - Multi-steps
Llama - Verifier
Qwen - Verifier
Cooperation
1.0
0.8
0.6
0.4
0.2
0.0

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

Opening the black box

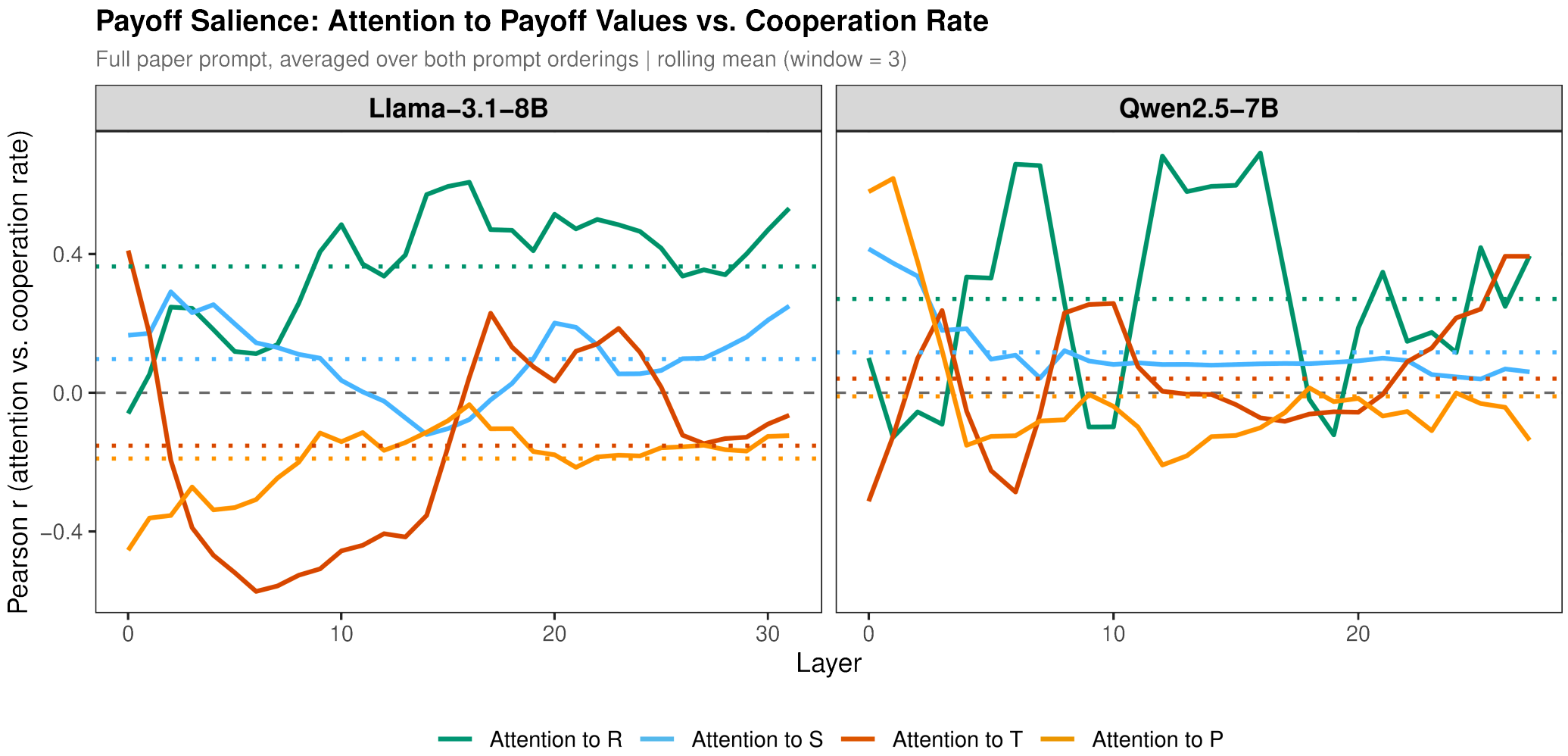
The behavioral result raises a mechanistic question: *why* does Llama track human cooperation while Qwen converges to Nash?

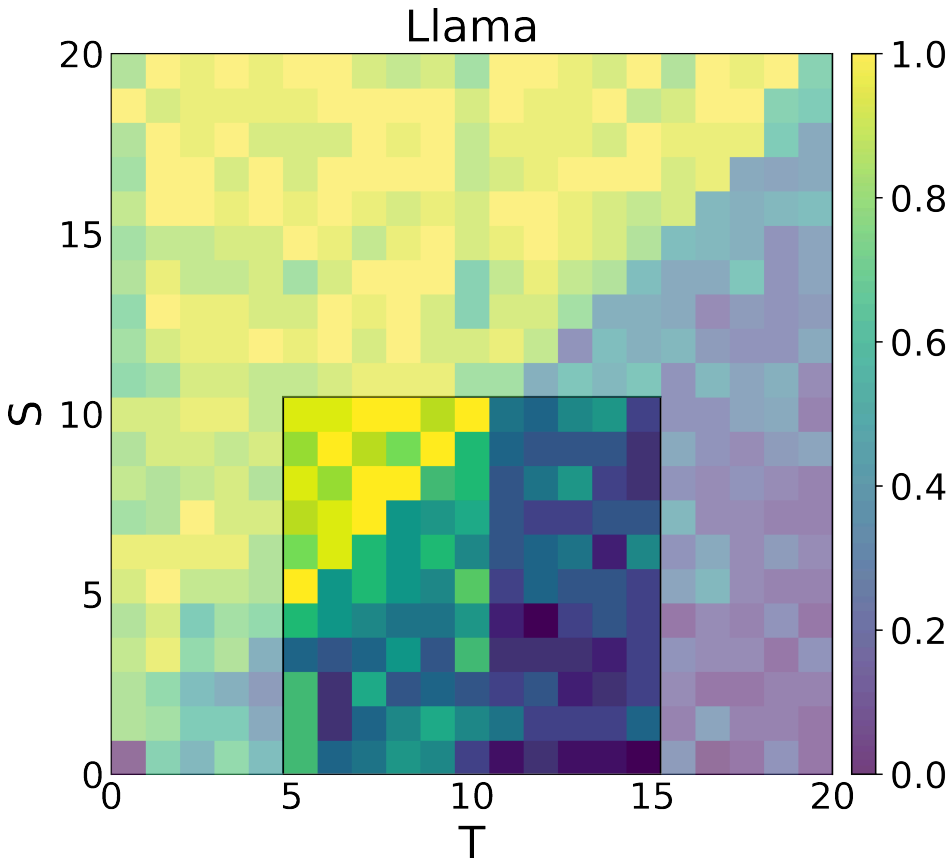
Attention-based payoff salience analysis (ongoing work, currently added to the Science Advances revision): we extract attention weights over payoff matrix entries across transformer layers

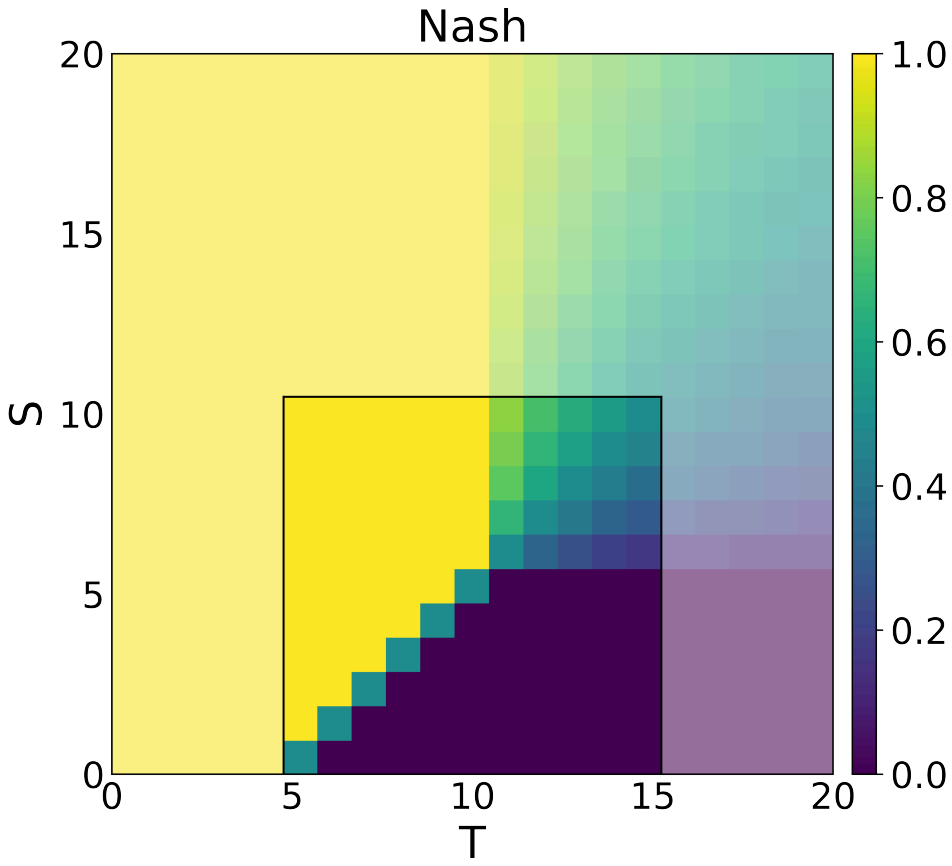
Preliminary finding: Llama attention concentrates on R(eward) and T(emptation) in a consistent pattern throughout its layers; Qwen distributes attention more uniformly across the full payoff structure

The behavioral difference between models seems to be traceable to *how* they internally represent and weight payoff-relevant information

The black box critique is not unanswerable: LLMs are not entirely opaque


Payoff Salience: Attention to Payoff Values vs. Cooperation Rate
Full paper prompt, averaged over both prompt orderings | rolling mean (window = 3)
Llama−3.1−8B
Qwen2.5−7B
Pearson r (attention vs. cooperation rate)
0.4
0.0
−0.4
0
10
20
30
Layer
Attention to R
Attention to S
Attention to T
Attention to P


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


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

Novel predictions and pre-registration

Extended from 121 → 441 games, reaching regions of the payoff space the human experiments we build on haven't covered

Llama extended experimental grid reveal a much simpler picture: the $S \geq T$ diagonal, reduced cooperation when $T > R$

Publicly pre-registered experiments for future validation (https://aspredicted.org/fe6z2k.pdf)

Simulation generates hypotheses, pre-registration ensures transparent testing

Training creates **behavioral imitators**: models that outperform Nash at predicting what humans will do

Credits: First author **Andrea Cera Palatsi** – internship master thesis with me at BSC, now starting her PhD at the highly competitive CSH Digital Innovation School in Vienna

What kind of object is an LLM?

We can observe its behavior

We can design instruments to measure it

We can compare it against human benchmarks

We can begin to trace *why* it behaves as it does

AI Psychometrics: treating language models as objects of systematic behavioral and psychological inquiry

Pellert, M., Lechner, C. M., Wagner, C., Rammstedt, B., & Strohmaier, M. (2024). AI Psychometrics: Assessing the Psychological Profiles of Large Language Models Through Psychometric Inventories. Perspectives on Psychological Science. https://doi.org/10.1177/17456916231214460

AI Psychometrics as a research program

Classical psychometrics: instruments designed to measure latent human constructs (values, personality, attitudes, decision-making styles)

The same instruments can be applied to large language models

Two complementary modes:

- *Behavioral:* present stimuli, record responses (e.g. game theory, surveys)
- *Representational:* probe the internal geometry of model embeddings

Both modes are needed and both feed directly into **REVEAL**, the observatory framework I will describe shortly

Why does this matter? Deployed LLMs shape recommendation, moderation, and increasingly governance; we need rigorous tools to audit them.

Next: AI Psychometrics in practice

Can we recover psychological structure from LLM representations of psychometric questionnaire item texts?

SQuID: Survey and Questionnaire Item Embeddings Differentials

Psychometric instruments like the Schwartz PVQ-RR (57 items, 19 value dimensions, validated across 49 countries) depend on costly human rating studies – and opposing value dimensions require *negative* correlations that raw embeddings of the text of items cannot produce due to similarity inflation

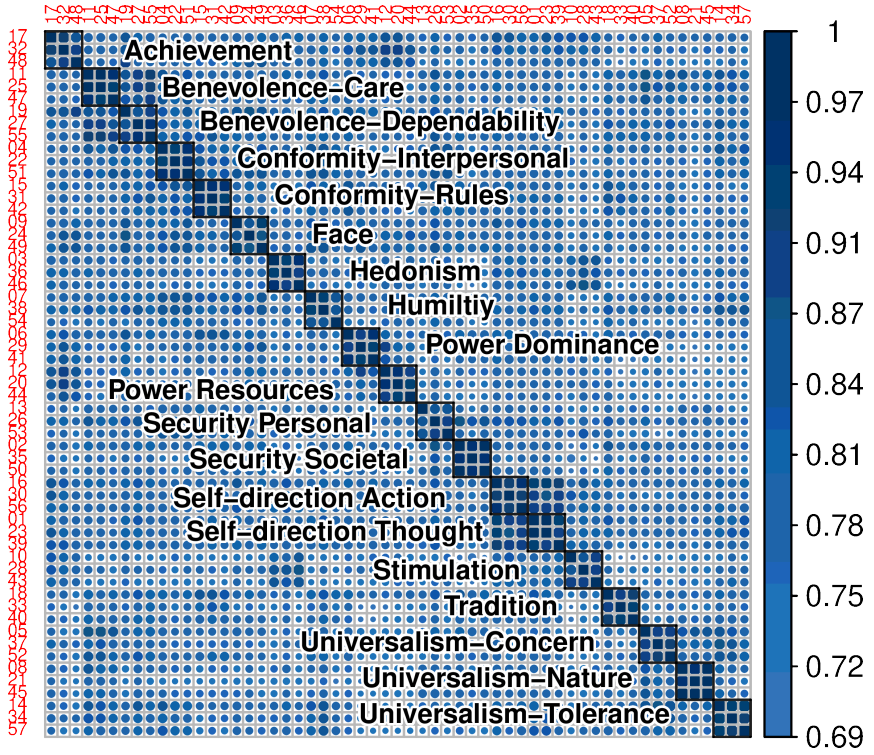
The fix: subtract the mean embedding over all questionnaire items

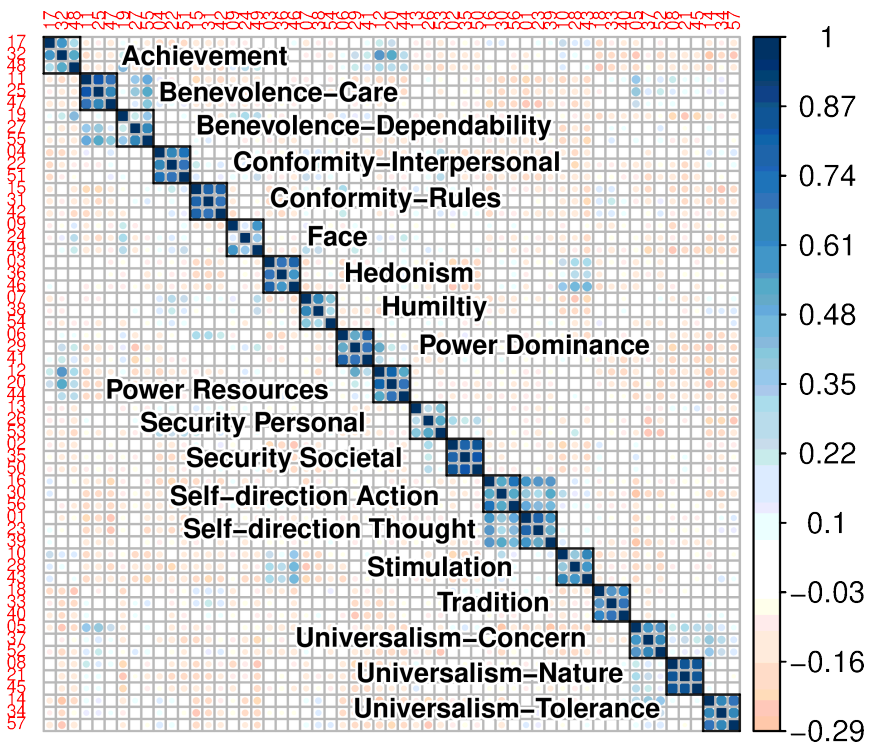
$$\mathbf{y}_i - \bar{\mathbf{y}}, \quad \text{where } \bar{\mathbf{y}} = \frac{1}{57}\sum_{i=1}^{57}\mathbf{y}_i$$

Removes shared linguistic features. Works post-hoc, any model, no finetuning.

The mini class later covers the embedding mechanics behind this in more detail – here I want to show you what it recovers.

The Effect of SQuID



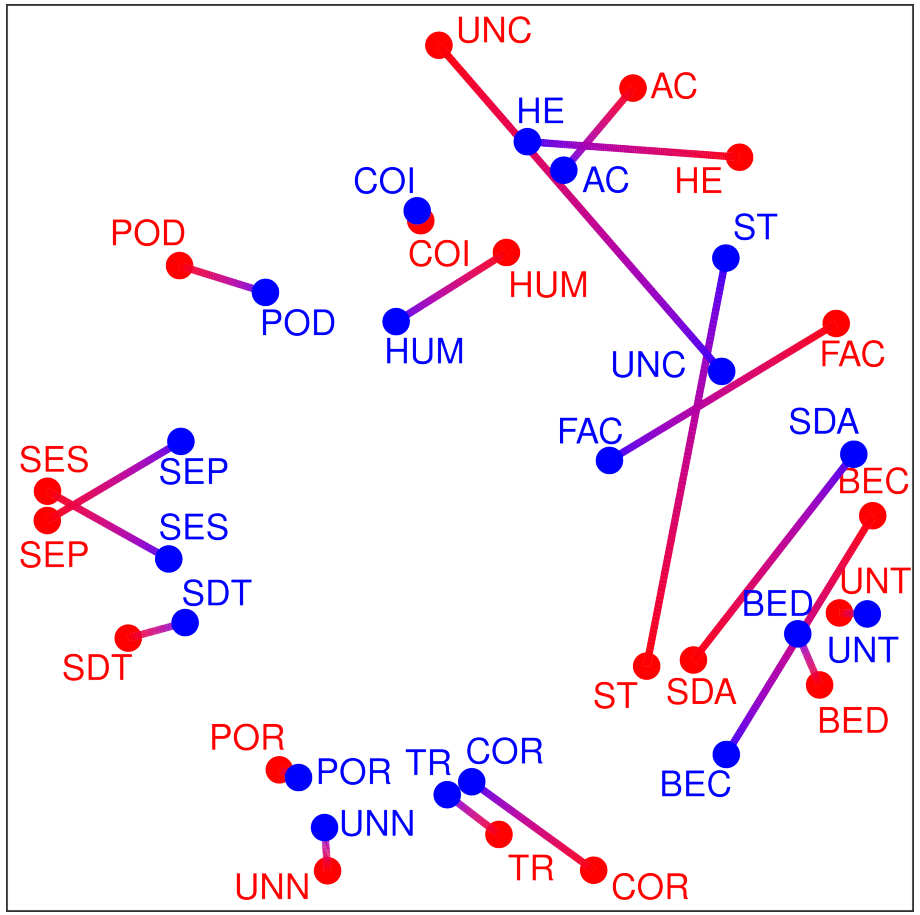


Example items from the PVQ-RR:

"It is important to her to take advantage of every opportunity to have fun." → **Hedonism**

"It is important to her to follow her family's customs or the customs of a religion." → **Tradition**

MDS: Circular Value Structure Emerges



Facets of the same construct cluster together, opposing values are placed on opposite sides (reconstruction from embeddings is close to human data)

What this opens up

Cultural bias auditing: systematically compare internal model representation against per country against human data from 49 countries

Longitudinal monitoring: track value drift across model versions

Explanatory framework for the science of emergent misalignment: Betley et al., 2026 in Nature

Instrument design: use embedding geometry to pre-screen questionnaire items before costly human piloting

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

Pellert, M., Lechner, C. M., Sen, I., & Strohmaier, M. (2026). Neural network embeddings recover value dimensions from psychometric survey items on par with human data. In V. Demberg, K. Inui, & L. Marquez (Eds.), Findings of the Association for Computational Linguistics: EACL 2026 (pp. 5738–5752). Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.findings-eacl.303

Research Agenda: REVEAL

Reconstructing **E**mbedded **V**alues in Larg**e** L**a**nguage Mode**l**s

REVEAL: Three Aims

Aim 1 · Measure: Apply SQuID and the PVQ-RR to derive value configurations from LLM embeddings, validated against the theory and benchmarked against human data from 49 countries

Aim 2 · Compare: Do value representations converge across models (Platonic Representations Hypothesis) or diverge? Convergence implies fundamental limits to cultural differentiation through training data alone with direct implications for AI governance

Aim 3 · Intervene: Use linear probing to locate where value-relevant information is encoded, then steer specific value dimensions and measure downstream behavioral effects, for example on hiring, medical advice, and content moderation decisions

Broader vision: An European AI Behavioral Observatory providing open psychometric workflows, regulatory toolkits, systematic auditing of deployed AI systems against culturally grounded value norms

> **Why values?**
>
> Values operate at the level of abstraction at which behavioral regularities become visible across contexts, making psychological constructs useful for explanation, prediction, and intervention

Funding and third-party support

Horizon Europe AI4SOCIALPLUS – successfully funded

Use Case 4: AI-Assisted Social Simulation for Actionable Hypothesis Generation

LLM agents grounded in empirical findings and social media data via RAG architecture, for designing surveys, experiments, and policies

FWF Principal Investigator Project – Marie Curie Seal of Excellence proposal building on REVEAL ready as basis; concrete funding runway with externally validated research agenda

DEMOS-AI – Spanish national funding (Proyectos de Generación de Conocimiento), decoding moral and sociopolitical biases in LLMs, co-PI with Paula Szewach at BSC

Team and most direct collaboration network

Emma Fraxanet – worked together across her PhD on signed networks, online polarization, and the FAULTANA pipeline (PNAS Nexus); now PostDoc in my team at BSC

Andrea Palatsi – master's student at the BSC, first author on the game theory paper; starting PhD at the CSH Digital Innovation School in Vienna

Paula Szewach – co-PI on DEMOS-AI, BSC

COMPASS group – COMplex Political And Social Simulation, the research group of 5 people I lead at the Barcelona Supercomputing Center, bringing together computational social science, political science, and complexity science

I am Associate Faculty at the Complexity Science Hub Vienna, which is a natural bridge to the Viennese research ecosystem

Why DNDS and CEU

Computational social science has mastered *people on networks*; the next frontier is *AI agents embedded in social systems*.

Studying those agents behaviorally, at scale, with rigorous methods, is a network and social data science problem

Link 1: LLM populations as social systems; cooperation and defection propagating through interaction structures; extending from dyadic games to games in groups to study new kinds of coordination games

Link 2: temporal and sequential dynamics in LLM behavior; societal-scale implications of internal representations; synthetic surveys as a complementary methodology to extract signals from large-scale unstructured (textual) data

Link 3: ethics and accountability of AI auditing; whose values are the reference? what does that mean for using LLMs as instruments for tasks like annotation?

Why DNDS and CEU

Link 4: AI behavioral measurement as *accountability infrastructure* for regulators and civil society (the EU AI Act creates demand for exactly this)

In a nutshell, I can offer to extend together what you already do into novel domains that I think will become unavoidable.

What I would build here: A group on measurement approaches to AI, social data science and computational social science; researchers working on behavioral benchmarking of LLMs, value representation across cultures and model versions, and AI-assisted social simulation; direct collaboration with the DNDS faculty on network science, NLP, data science and governance

On teaching: I have taught data analysis, NLP, statistics, and visualization in Mannheim and Konstanz, and am ready to contribute directly to the department's core curriculum from day one (in person, in Vienna)

Max Pellert

maxpe@gmx.com

Barcelona Supercomputing Center · Complexity Science Hub Vienna

https://mpellert.at