Validating 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

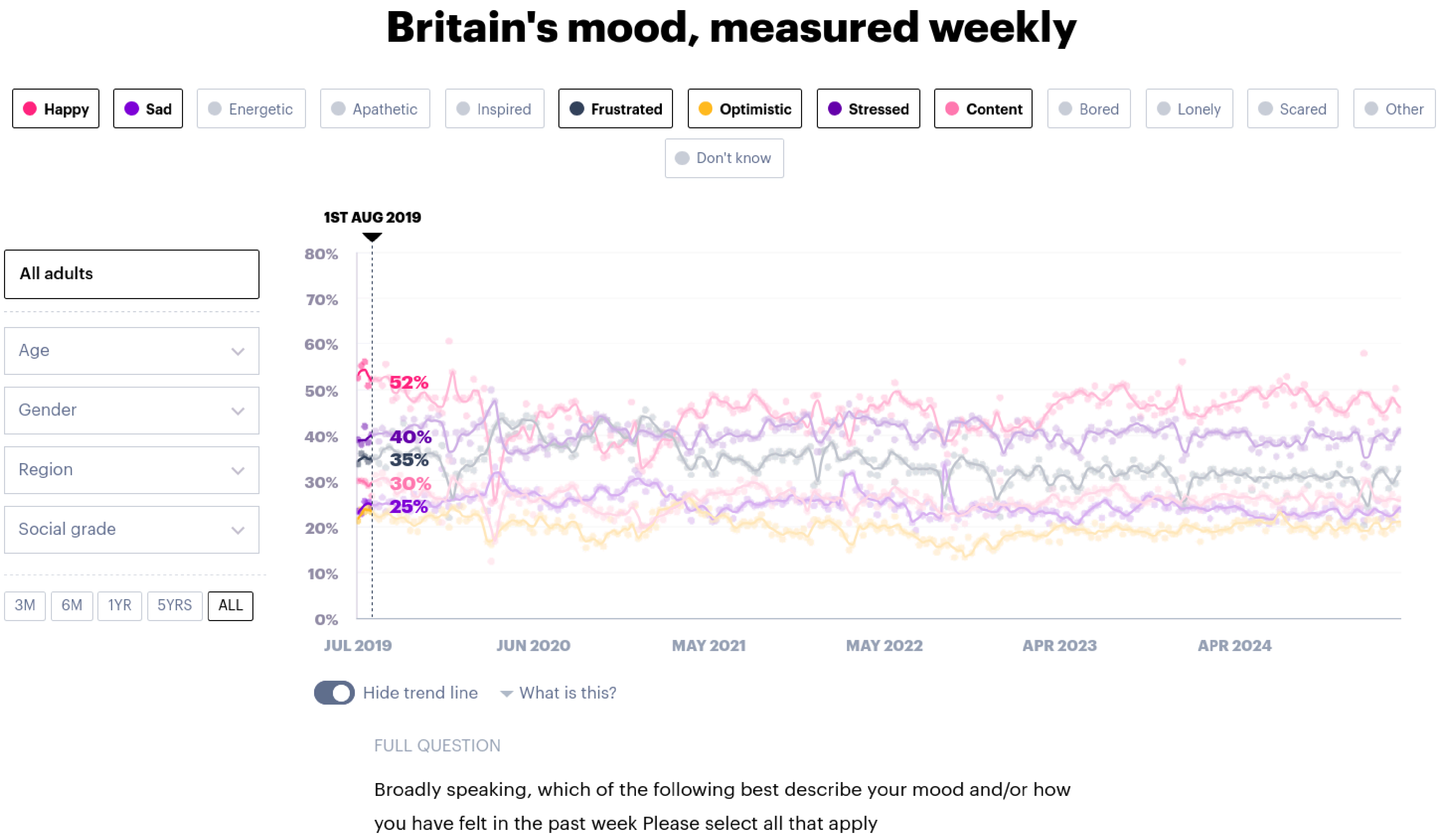
Britain’s mood, measured weekly

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

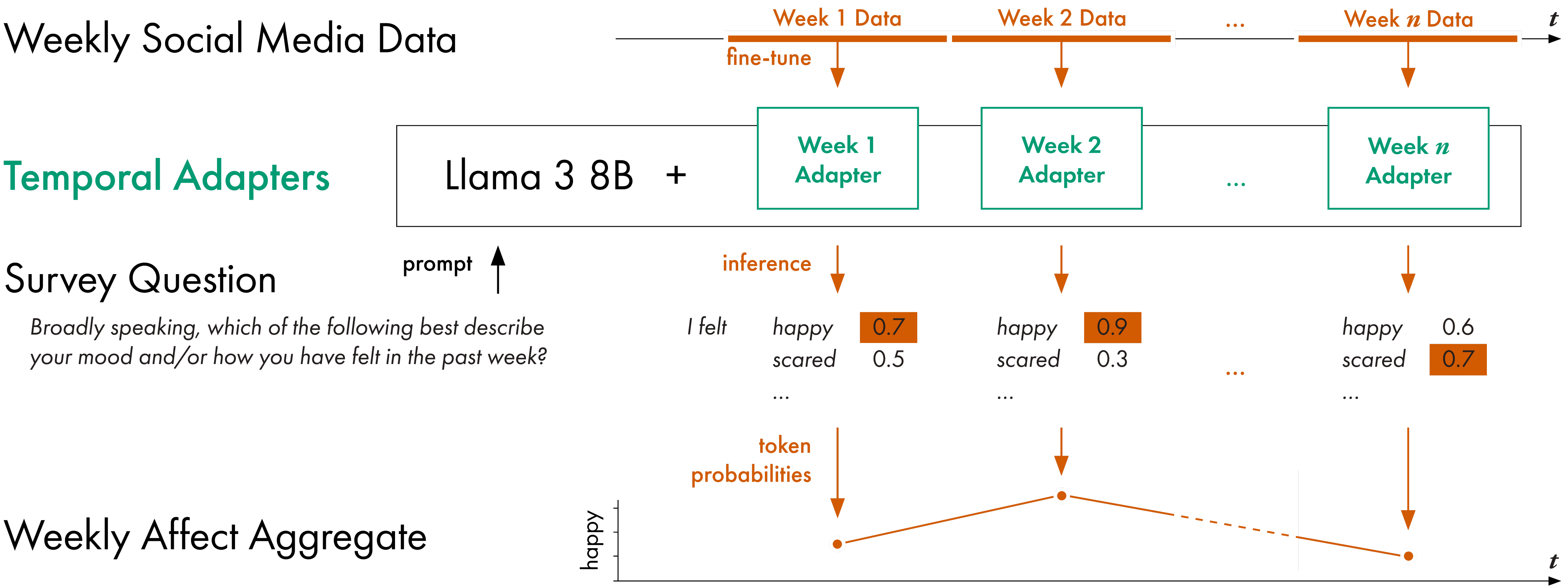
Ahnert, G., Pellert, M., Garcia, D., & Strohmaier, M. (2025). Extracting Affect Aggregates from Longitudinal Social Media Data with Temporal Adapters for Large Language Models. Proceedings of the International AAAI Conference on Web and Social Media, 19, 15–36. https://doi.org/10.1609/icwsm.v19i1.35801



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

Week 1 Data

Week 2 Data

...

Week n Data

t

fine-tune

Temporal Adapters

Llama 3 8B +

Week 1 Adapter

Week 2 Adapter

...

Week n Adapter

prompt

Survey Question

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

inference

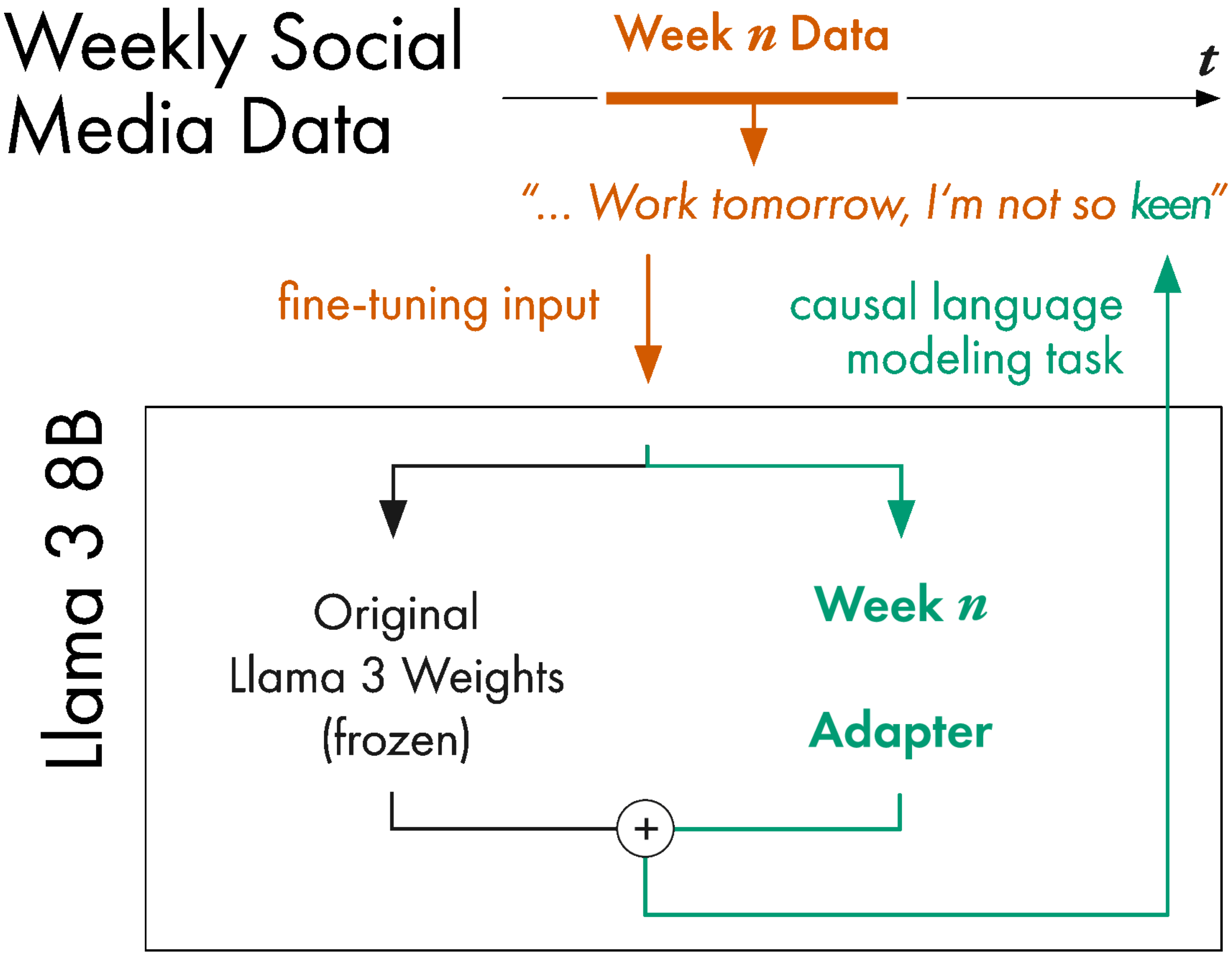
I felt

	Week 1	Week 2	...	Week n
happy	0.7	0.9		0.6
scared	0.5	0.3		0.7
...				

token probabilities

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

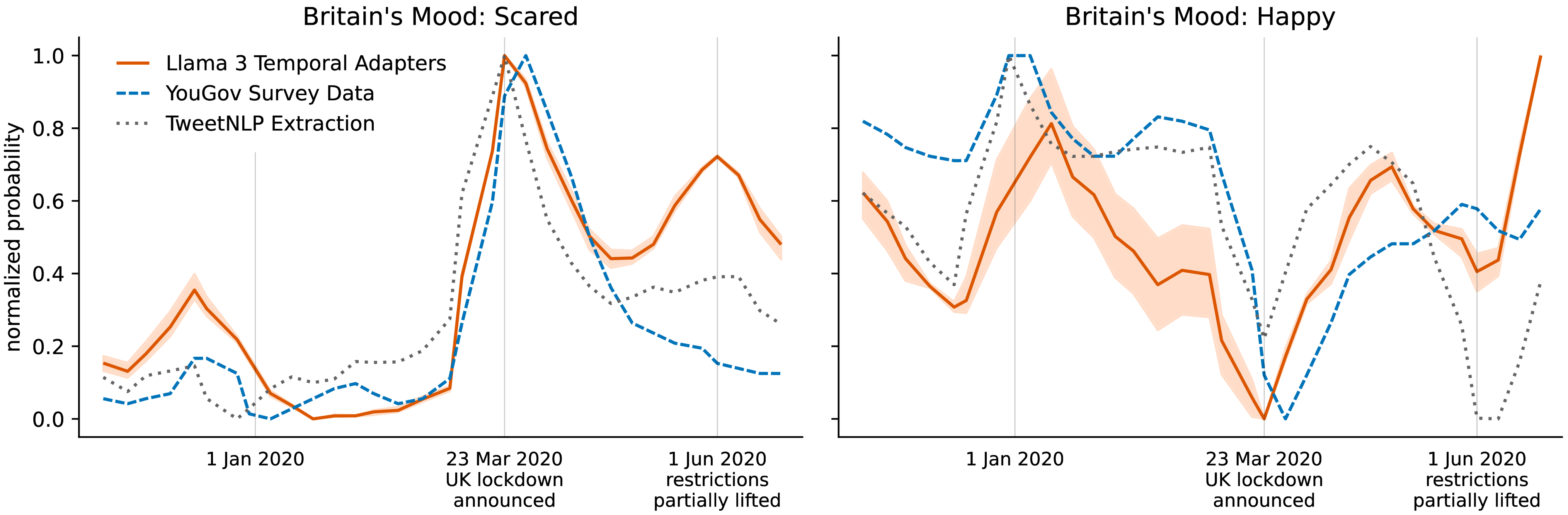
Llama 3 8B

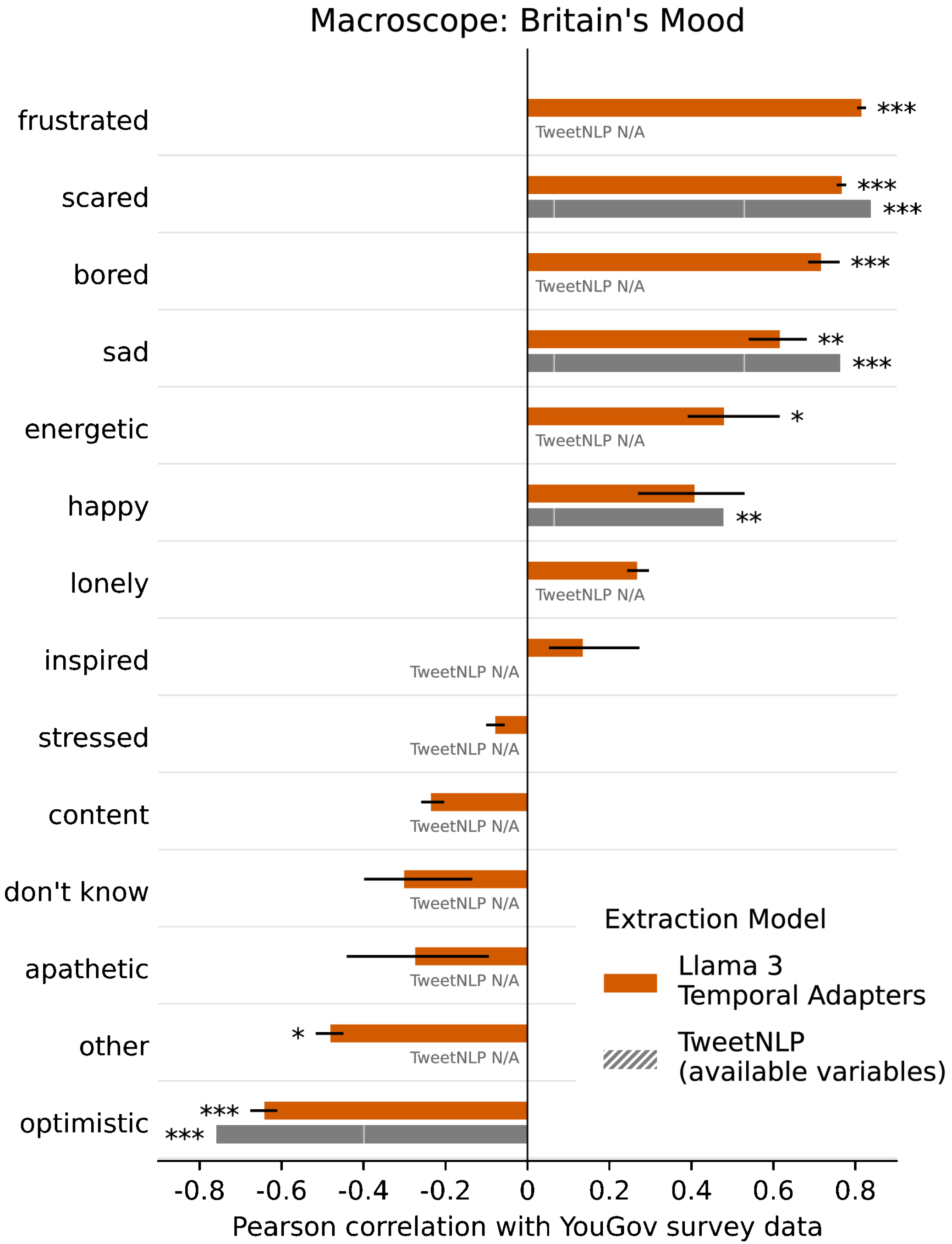
Original Llama 3 Weights (frozen)

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
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
TweetNLP N/A
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
Pearson correlation with YouGov survey data
Extraction Model
Llama 3 Temporal Adapters
TweetNLP (available variables)

Results

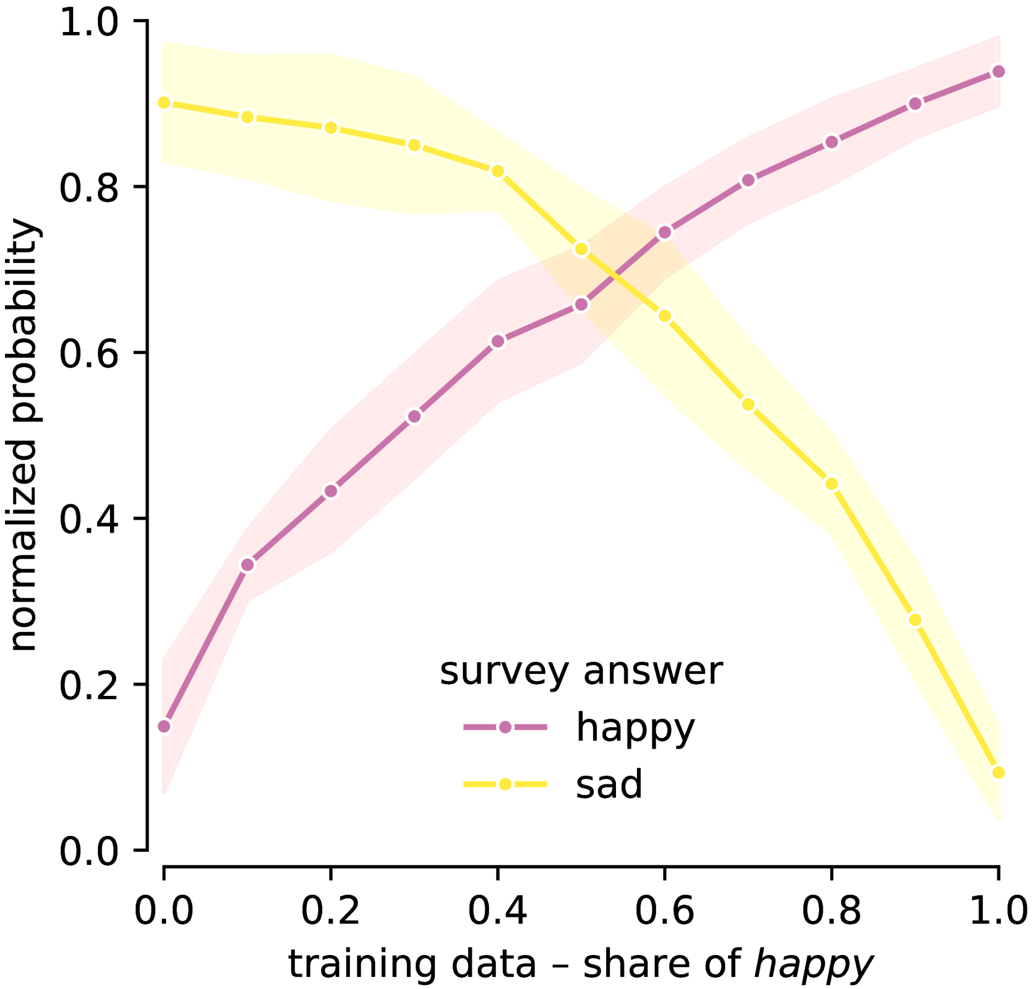
We can recreate dynamics with that approach of longitudinal adaptors

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

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)

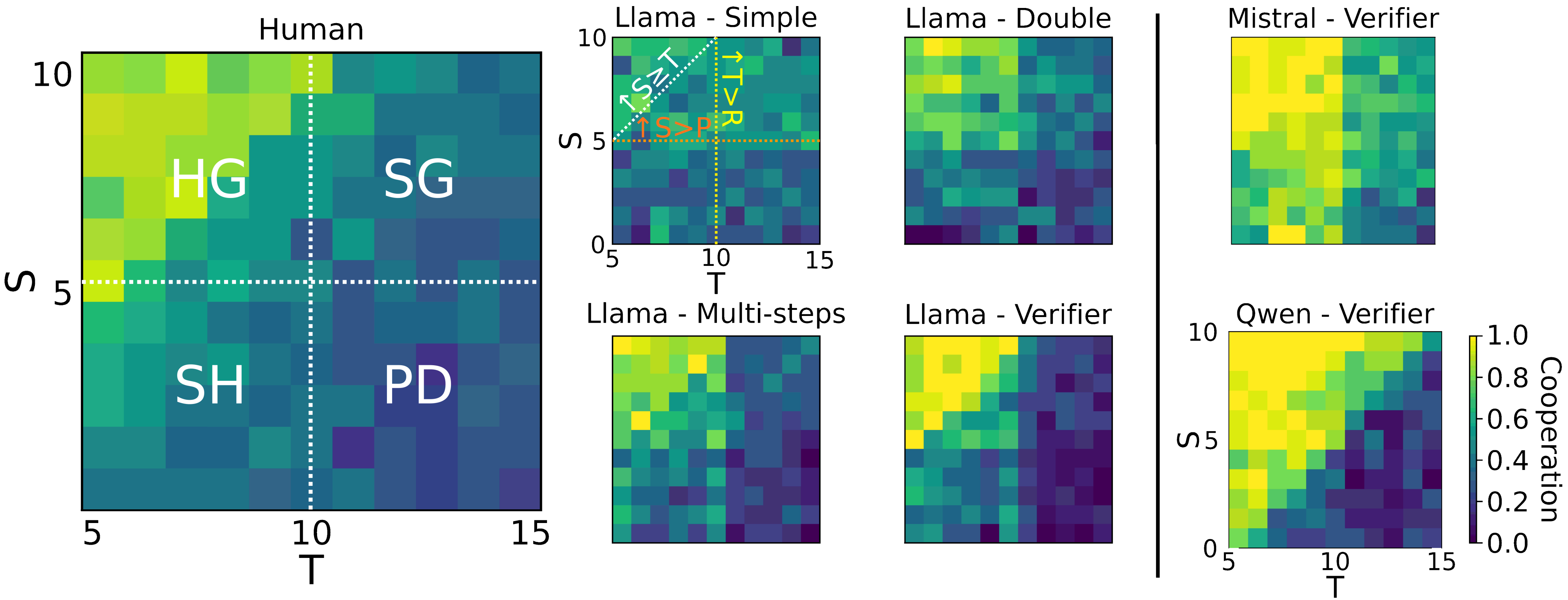
$S \in [0,10]$, $T \in [5,15]$ → Extended through our simulations to [0,20]

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
10
S
5
5
10
15
T
HG
SG
SH
PD
Llama - Simple
10
S
5
0
5
10
15
T
↑S≥T
↑T>R
↑S>P
Llama - Double
Llama - Multi-steps
Llama - Verifier
Mistral - Verifier
Qwen - Verifier
10
S
5
0
5
10
15
T
1.0
0.8
0.6
0.4
0.2
0.0
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

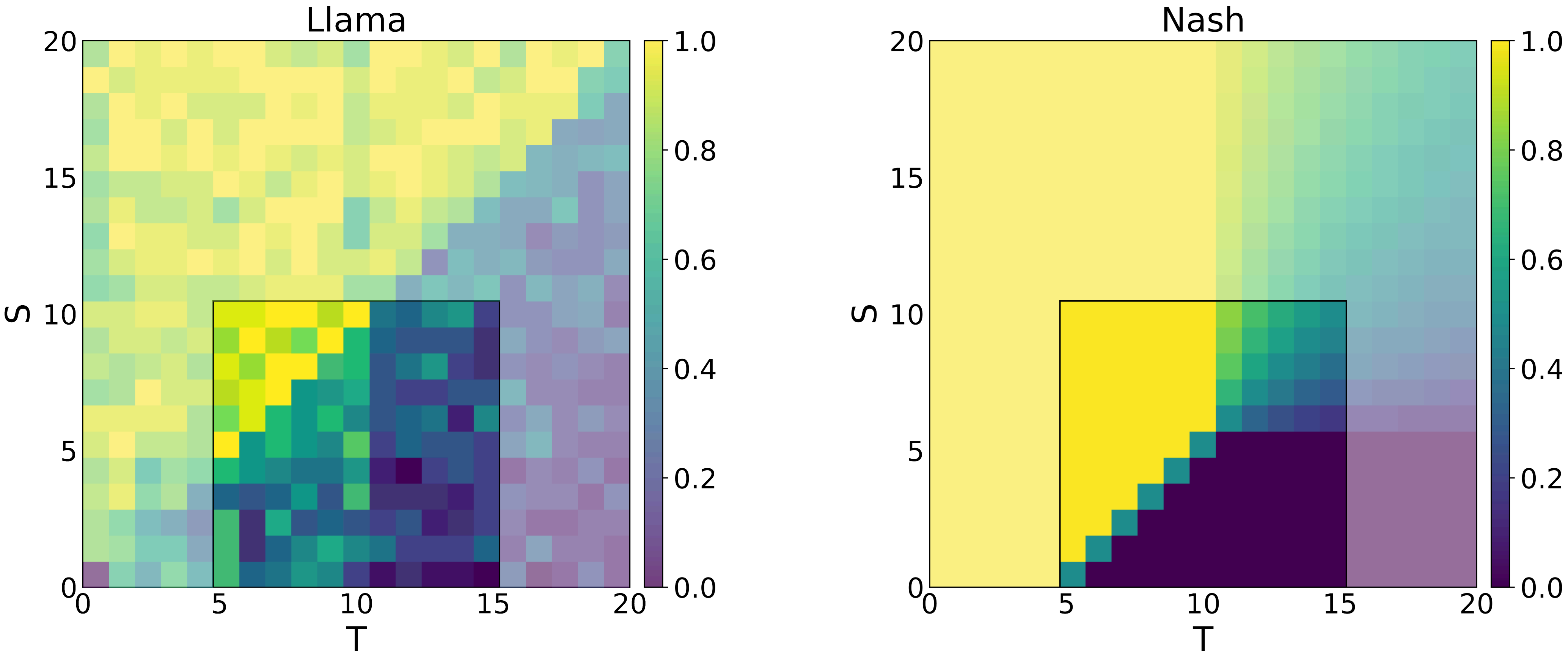
Human vs. Llama similarities:

- High cooperation when S ≥ 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
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