CLARIN.SI

Can We Trust AI as a Scientific Instrument for the Humanities?

Nikola Ljubešić
Jožef Stefan Institute, Ljubljana, Slovenija

12 November 2026, HumAI Seminar Series
Gothenburg, Sweden / Online

Jožef Stefan Institute, Ljubljana, Slovenia







CLARIN.SI

Wide collaboration network

CLARIN.SI
Slovenska raziskovalna infrastruktura
za jezikovne vire in tehnologije



cjvt

CLASSLA
K-centre





ReLDI



South-Slavic language space
CLASSLA K-Centre (CLARIN ERIC)



CLARIN.SI

Where I come from

Natural Language Processing / Computational Linguistics

2010s – Large web corpora for South Slavic languages (hrWaC, slWaC...)
billions of words in size, great for language research, but also...

2017 the transformer model 2018 BERT pre-trained encoder model

2019 CamemBERT 🇫🇷 AlBERTo 🇮🇹 BERTje 🇳🇱

Ljubešić and Lauc (2021) BERTić 🇭🇷🇧🇦🇲🇪🇷🇸

Ljubešić et al. (2024) continued pre-training XLM-R-BERTić 🇭🇷🇧🇦🇲🇪🇷🇸

Around that time hell was breaking loose...

CLARIN.SI

How I cope (deep from the bushes)

1. Benchmarks in language (and dialect) understanding

   (Can we trust the new scientific instrument?)
2. Going beyond text into speech (and other modalities?)

   (Performing measurements on human signal at scale)
3. Fundamental research!

   (From AI measurements to scientific findings)

Benchmarks

(Can we trust AI as a scientific instrument)

CLARIN.SI

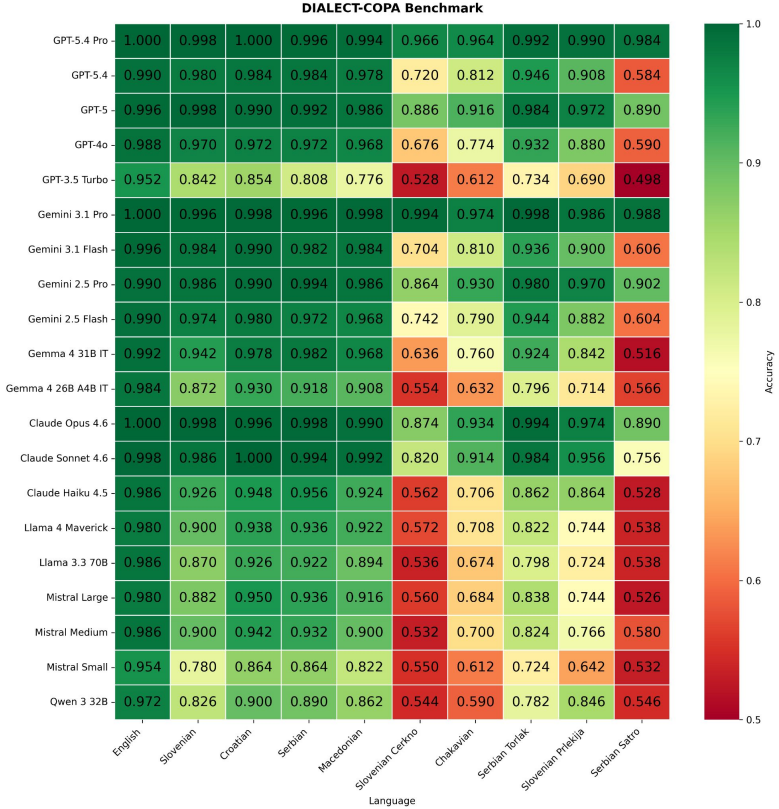
Language understanding benchmarks

Standard-language close-ended NLU benchmarks are “cooked”

Dialects are still challenging, frontier is saturating, but open-weight are not

Šatra – syllable flipping argo (secret language used by the early-20-century criminals in northern Croatia and Serbia)

“bačkizagre stuhpa šeja”



CLARIN.SI

Benchmarks through wide collaboration!

We need language- and culture-specific benchmarks

Translating a benchmark will not do!

Benchmarks get saturated fast!

Good example of efficient collab: MRL 2025 GlobalPIQA benchmark

Gemini 3 official eval

NeurIPS paper! (1/380 of it is mine)

Benchmark	Description		Gemini 3 Pro	Gemini 2.5 Pro	Claude Sonnet 4.5	GPT-5.1
Humanity's Last Exam	Academic reasoning	No tools With search and code execution	**37.5%** **45.8%**	21.6% —	13.7% —	26.5% —
ARC-AGI-2	Visual reasoning puzzles	ARC Prize Verified	**31.1%**	4.9%	13.6%	17.6%
GPQA Diamond	Scientific knowledge	No tools	**91.9%**	86.4%	83.4%	88.1%
AIME 2025	Mathematics	No tools With code execution	**95.0%** **100%**	88.0% —	87.0% **100%**	94.0% —
MathArena Apex	Challenging Math Contest problems		**23.4%**	0.5%	1.6%	1.0%
MMMU-Pro	Multimodal understanding and reasoning		**81.0%**	68.0%	68.0%	76.0%
ScreenSpot-Pro	Screen understanding		**72.7%**	11.4%	36.2%	3.5%
CharXiv Reasoning	Information synthesis from complex charts		**81.4%**	69.6%	68.5%	69.5%
OmniDocBench 1.5	OCR	Overall Edit Distance, lower is better	**0.115**	0.145	0.145	0.147
Video-MMMU	Knowledge acquisition from videos		**87.6%**	83.6%	77.8%	80.4%
LiveCodeBench Pro	Competitive coding problems from Codeforces, ICPC, and IOI	Elo Rating, higher is better	**2,439**	1,775	1,418	2,243
Terminal-Bench 2.0	Agentic terminal coding	Terminus-2 agent	**54.2%**	32.6%	42.8%	47.6%
SWE-Bench Verified	Agentic coding	Single attempt	76.2%	59.6%	**77.2%**	76.3%
τ2-bench	Agentic tool use		**85.4%**	54.9%	84.7%	80.2%
Vending-Bench 2	Long-horizon agentic tasks	Net worth (mean), higher is better	**$5,478.16**	$573.64	$3,838.74	$1,473.43
FACTS Benchmark Suite	Held out internal grounding, parametric, MM, and search retrieval benchmarks		**70.5%**	63.4%	50.4%	50.8%
SimpleQA Verified	Parametric knowledge		**72.1%**	54.5%	29.3%	34.9%
MMMLU	Multilingual Q&A		**91.8%**	89.5%	89.1%	91.0%
Global PIQA	Commonsense reasoning across 100 Languages and Cultures		**93.4%**	91.5%	90.1%	90.9%
MRCR v2 (8-needle)	Long context performance	128k (average) 1M (pointwise)	**77.0%** **26.3%**	58.0% 16.4%	47.1% not supported	61.6% not supported

CLARIN.SI

Speech processing

(Measuring human signal at scale)

CLARIN.SI

Why speech processing

Text is a great knowledge preserving technology

But human communication is all about speech (+ nonverbal)!

Speech processing used to be extremely hard (MFCC / filterbank), drastic improvements with transformers

VERY important announcement:
Speech is NOT audible text!
(Moore, 2025)

ParlaSpeech

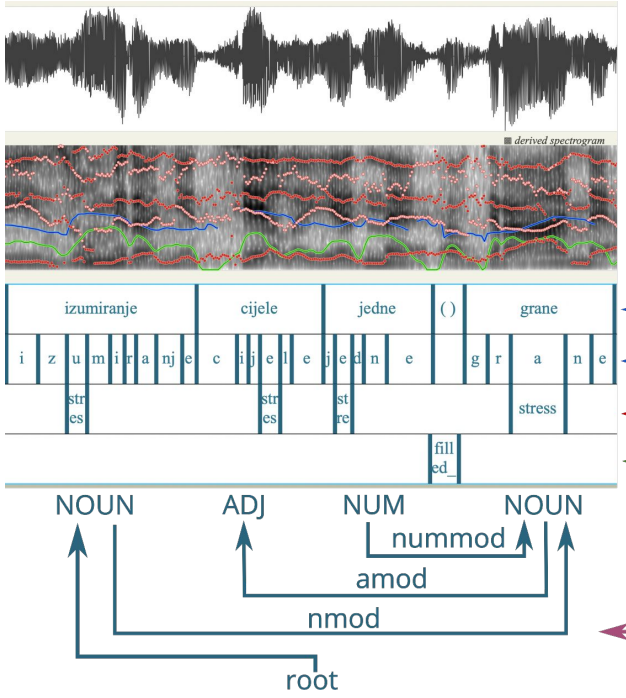
10k hours of speech from 5 Slavic languages

Automatic enrichments

Multilingual speech and text encoders

Strong cross-lingual performance

Strong research potential: interaction of so many layers



Word alignment

Grapheme alignment

Primary (word) stress

Filled pauses

Sentiment: "Negative"

Topic: "Agriculture"

Linguistic annotations

Rich metadata:

Name: "Ružica Vukovac"
Gender: "Female"
Birth year: "1975"
Party: "MOST"
Status: "Opposition"
...

CLARIN.SI

The encoders need fine-tuning data...

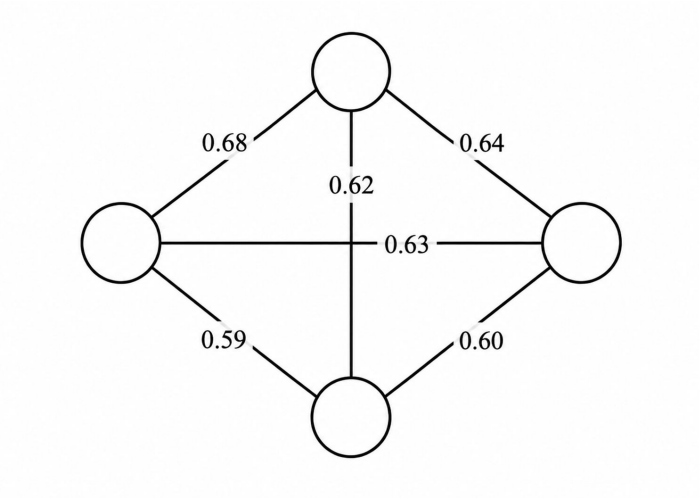
We need supervision (labels) for fine-tuning speech and text encoders

We also need evidence that the operationalization is robust

Inter-rater agreement (Krippendorff's α)
between three humans and one machine.

Guess which one is the machine...

Teacher frontier model,
student small open encoder,
ensuring 1. scaling,
2. reproducibility, 3. accessibility.



(Kuzman and Ljubešić, 2025)

CLARIN.SI

Dialectal speech recognition

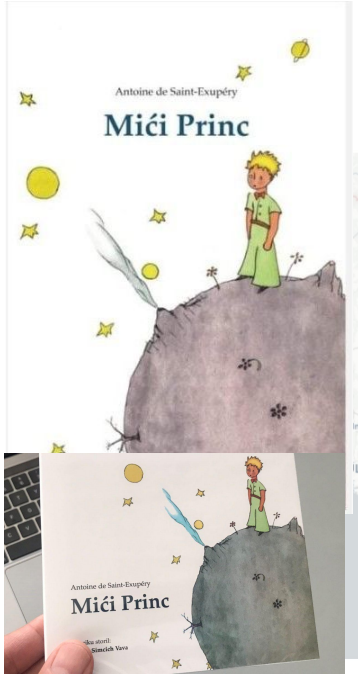
Chakavian dialect

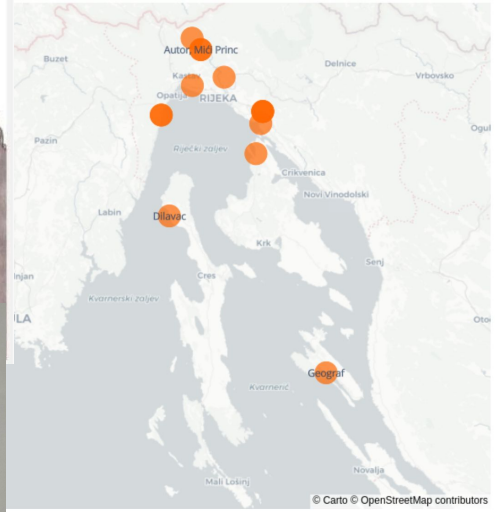
Ljubešić et al. (2024)

A 70-minute dataset can already halve the word-error rate (to 15% WER)

Micro-dialects - every village speaks differently

Again – speech is NOT audible text, so ASR is just an accessibility layer





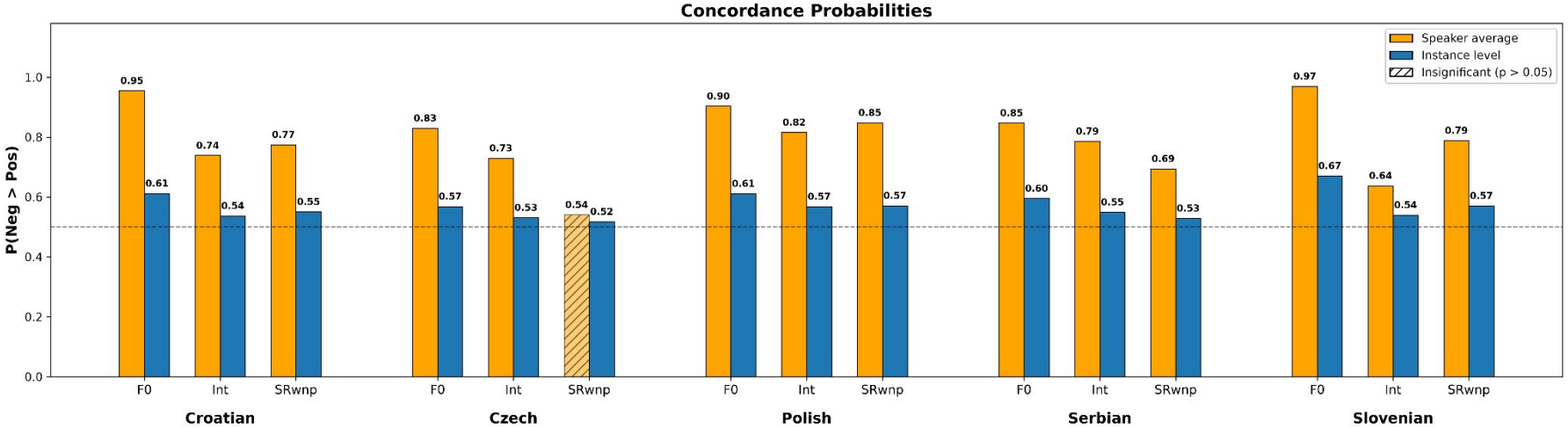
CLARIN.SI

Language research
(AI as scientific instrument)

PolSci / Rhetorics research on ParlaSpeech

Ljubešić et al. (2026)

Politicians, when negative, 1. raise voice, 2. talk louder, 3. talk faster



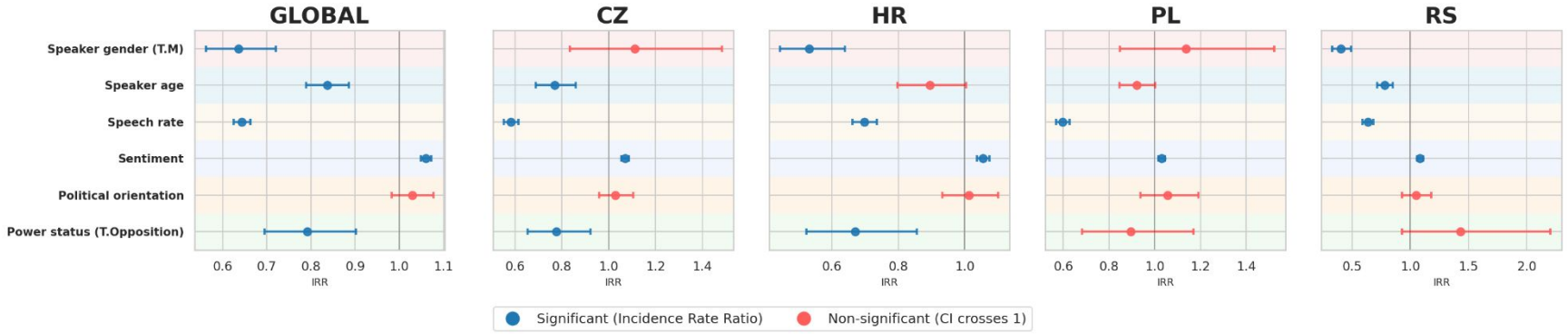
CLARIN.SI

Linguistic / PolSci research on ParlaSpeech

Porupski et al. (2026)

Predictors of filled pause (*eeem*) use

Less present in faster, negative speech, across opposition (except Serbia)



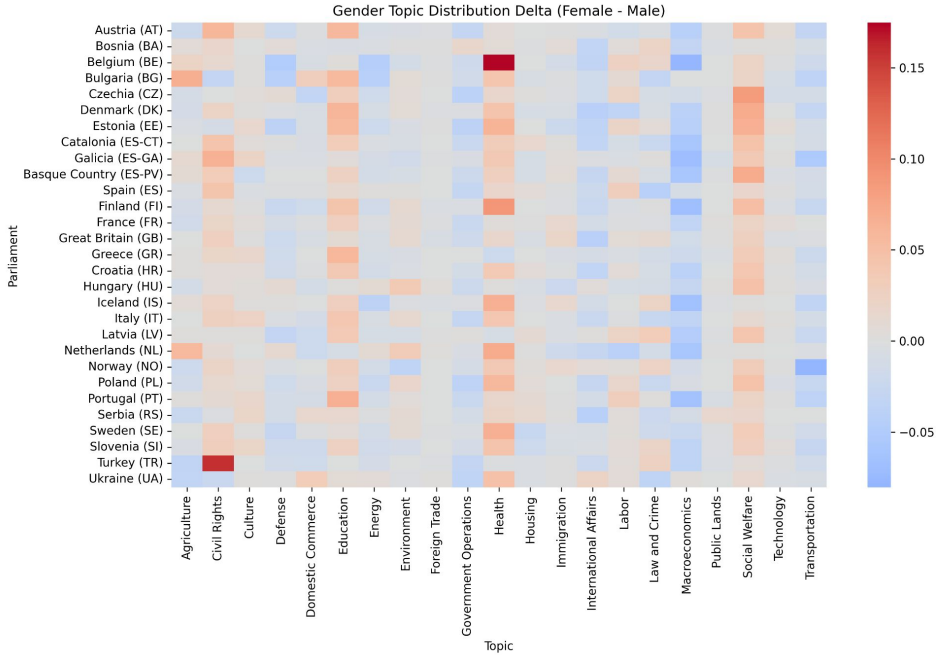
CLARIN.SI

PolSci research on ParlaCAP

Gender trends across topics in 28 European parliaments

Women discuss consistently more Health, Education, Social Welfare, Civil Rights

Men discuss more Macroeconomics, Transportation



Kuzman Pungeršek et al. (2026)

CLARIN.SI

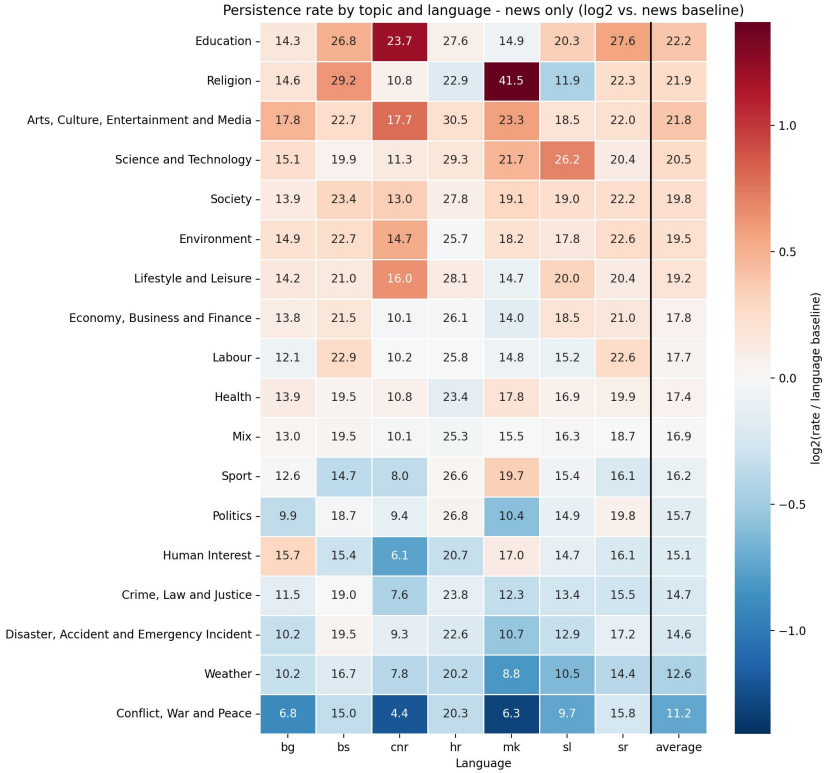
What persists on the Web

CLASSLA-web – 2022 (11B words) and 2024 (17B words) crawls of South Slavic web

Only 20% content available two years later

Event-based topics have a faster turnover than more static topics

News, promotion, instruction have a faster turnover than literature, opinion, information



CLARIN.SI

The road ahead

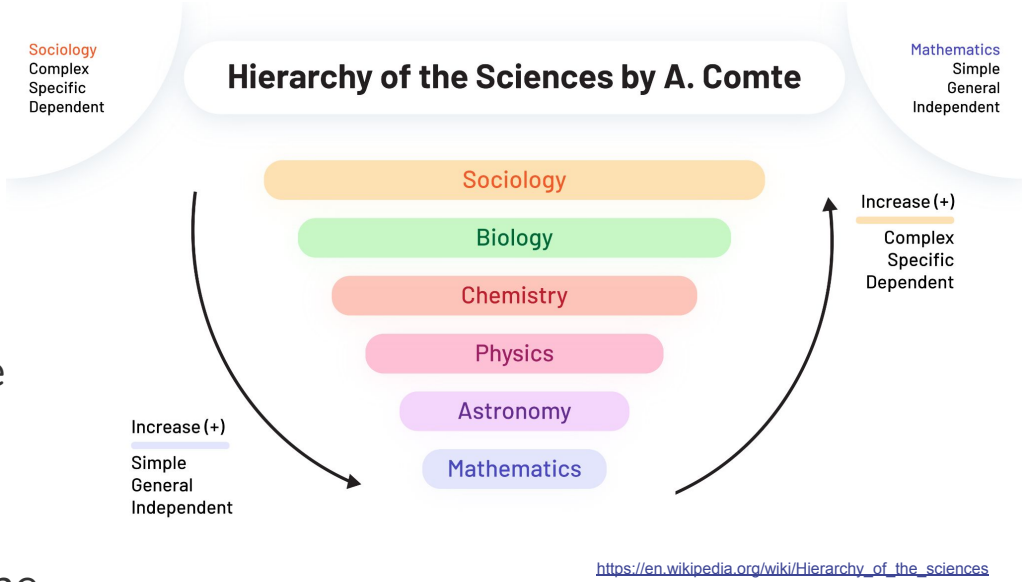
AI for Science

"Mathematics is crumbling" due to verifiability

Theoretical physics is being attacked as well

Up the "natural science path" we need robotics / automation

What is with SSH? Can we have verifiability at all? Will SSH be the last bastion of "human-dominated science"?



https://en.wikipedia.org/wiki/Hierarchy_of_the_sciences

It should not be!

CLARIN.SI

AI for SSH

Verifiability in SSH, can we reach it?

- Silicon sampling? No. (Remember dialectal results?)
- More complex simulations? Agentic swarms? Not any time soon.

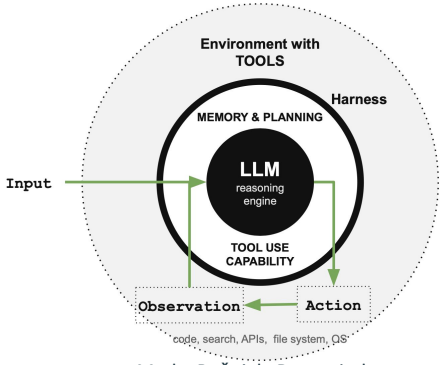
Let's take it slow(er) (as the "soft", most complex! sciences require)

- Genuine human-generated data at scale
- Operationalization with inter-rater agreement as proof-of-quality
- Encoder models as highly scalable and robust scientific instruments
- Validated, reproducible empirical methods (old or new)
- Replication across populations, (LLM-supported) multiverse analyses
- LLMs as student-model teachers and co-researchers!

LLM as a co-researcher

Agentic LLMs (tool calling, web search, coding, data wrangling) should already be (and is?) part of your research team!

- Research questions, study design & related work
- Data exploration & analysis
- Coding, statistics & visualisation
- Interpretation & robustness checks
- Writing & critical revision



Matko Bošnjak, Deepmind

My recent experience: surprisingly close to a post-doc collaborator!

But: Every link in the chain needs to be verified (multiple times).
By human and / or machine.

Would not have been possible without…



Taja Kuzman Pungeršek



Peter Rupnik



Ivan Porupski



Vuk Dinić



Michal Mochtak
Radboud University



Danijel Koržinek
PJIIT

… and MANY other collaborators!

CLARIN.SI

Thank you!

Questions? Comments?

nikola.ljubesic@ijs.si

https://nljubesi.github.io







CLARIN.SI