The importance of local LLMs and supporting HPC infrastructure for social sciences

Aleksandar Tomašević, University of Novi Sad

EuroCC4SEE Workshop in Belgrade, May 2025

Ursula von der Leyen
President of the European Commission

Dear Professor,

We have stumbled upon €2 billion of unallocated funds in the EU budget. We would like to grant these funds to your institution. Could you please reply with a quick budget proposal outlining how you would use these funds?

Warm regards,
Ursula von der Leyen
President of the European Commission

Ursula von der Leyen
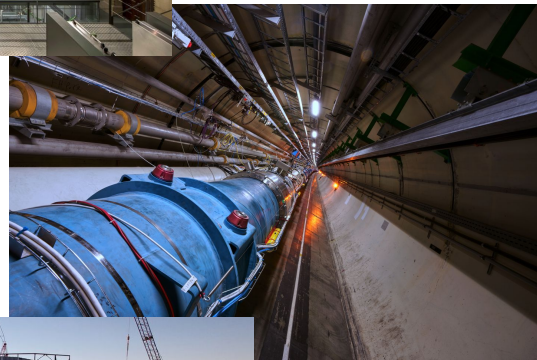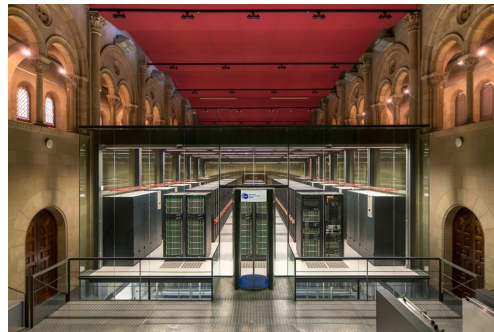President of the European Commission

Dear Professor,

We have stumbled upon €2 billion of unallocated funds in the EU budget. We would like to grant these funds to your institution. Could you please reply with a quick budget proposal outlining how you would use these funds?

Warm regards,
Ursula von der Leyen
President of the European Commission

Social Sciences: Lack of Scalable Research Infrastructure

Example: European Social Survey (ESS ERIC) has an annual budget

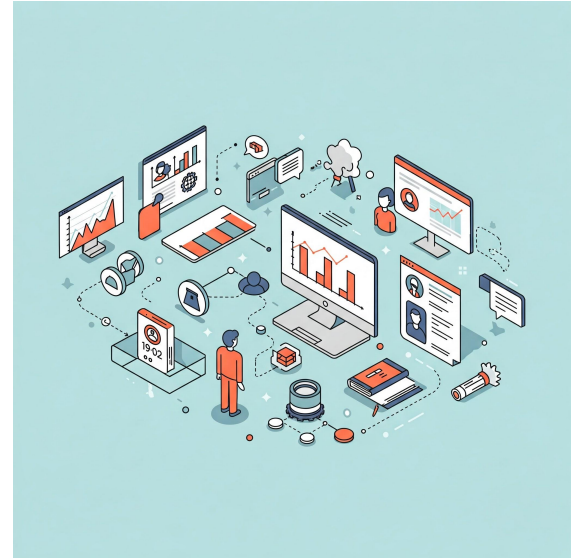~ <u>3 million €</u> (30+ countries)

What would happen if we increase that to <u>300 million €</u>?

Computational Social Science (CSS)

The use of computational methods to analyze massive digital datasets and model social phenomena

CSS requires **HPC infrastructure**

Computational Social Science (CSS)

MADOC: Multi-Platform Aggregated Dataset of Online Communities

Mitrovic Dankulov, Marija (Researcher)[1] iD; Tomašević, Aleksandar (Researcher)[2] iD;

Maletic, Slobodan (Researcher)[3] iD; Andjelkovic, Miroslav (Researcher)[3] iD;

Vranic, Ana (Researcher)[1, 4] iD; Cvetkovic, Darja (Researcher)[1] iD;

Stupovski, Boris (Researcher)[1] iD; Vudragovic, Dusan (Researcher)[1] iD;

Major, Sara (Researcher)[2] iD; Bogojević, Aleksandar (Researcher)[1] iD

Dataset Scale

- 18.9 million posts
- 236 million comments
- 23.1 million unique users across all platforms

IP8

CTRUST #7416 Prizma

Science Fund
of the Republic of Serbia

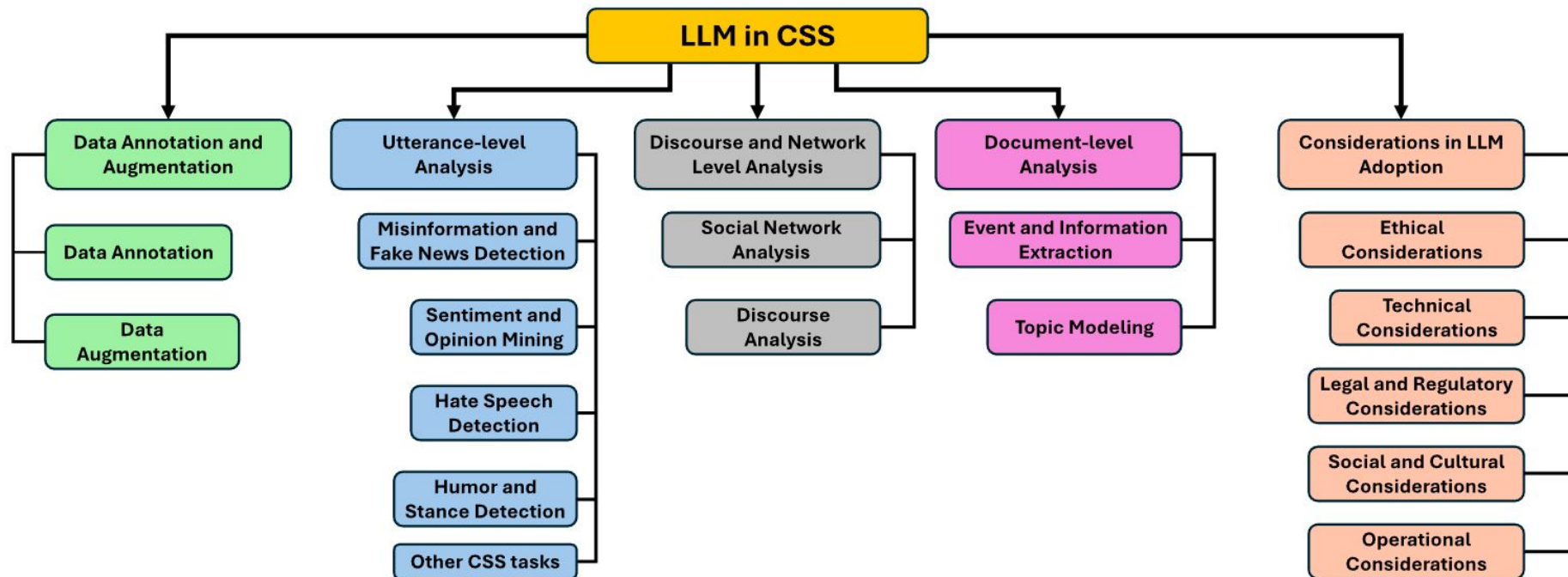# Computational Social Science (CSS)

With advancement of generative AI models, CSS offers social sciences an opportunity to catch up with the development and maintenance of HPC infrastructure.

Importance of local LLMs and HPC infrastructure

- **Reproducibility:** Ensuring verifiable research findings

- **Privacy:** Protecting sensitive data in social science research

- **Ollama** framework

- Why Social Sciences Need **HPC Infrastructure**?

Applications of LLMs in Computational Social Science

Thapa, S., Shiwakoti, S., Shah, S. B., Adhikari, S., Veeramani, H., Nasim, M., & Naseem, U. (2025). Large language models (LLM) in computational social science: prospects, current state, and challenges. *Social Network Analysis and Mining*, *15*(1), 1–30. https://doi.org/10.1007/s13278-025-01428-9

Reproducibility: Three ways to use LLMs

1. Chat interface

Reproducibility: Three ways to use LLMs

1. Chat interface

System Prompts

See updates to the core system prompts on **Claude.ai** and the Claude **iOS** and **Android** apps.

Claude's web interface (**Claude.ai**) and mobile apps use a system prompt to provide up-to-date information, such as the current date, to Claude at the start of every conversation. We also use the system prompt to encourage certain behaviors, such as always providing code snippets in Markdown. We periodically update this prompt as we continue to improve Claude's responses. These system prompt updates do not apply to the Anthropic API. Updates between versions are bolded.

Claude 3.7 Sonnet

▸ Feb 24th, 2025

Reproducibility: Three ways to use LLMs

1. Chat interface

System Prompts

See updates to the core system prompts on **Claude.ai** and the Claude **iOS** and **Android** apps.

Claude's web interface (**Claude.ai**) and mobile apps use a system prompt to provide up-to-date information, such as the current date, to Claude at the start of every conversation. We also use the system prompt to encourage certain behaviors, such as always providing code snippets in Markdown. We periodically update this prompt as we continue to improve Claude's responses. These system prompt updates do not apply to the Anthropic API. Updates between versions are bolded.

April 10, 2025

Sunsetting GPT-4 in ChatGPT

Effective April 30, 2025, GPT-4 will be retired from ChatGPT and fully replaced by GPT-4o.

Reproducibility: Three ways to use LLMs

2. Commercial API call

```python
from openai import OpenAI
client = OpenAI()

response = client.responses.create(
    model="gpt-4 ",
    input="Write a one-sentence bedtime story about a unicorn."
)

print(response.output_text)
```

Reproducibility: Three ways to use LLMs

2. Commercial API call

Reproducibility? ❌

Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine

Towards Automating Text Annotation: A Case Study on Semantic Proximity Annotation using GPT-4

Can GPT-4 Identify Propaganda? Annotation and Detection of Propaganda Spans in News Articles

Reproducibility: Three ways to use LLMs

3. Running local models

AdamDanielKing commented on May 29, 2020 ...

Most of us can hardly dream of using the full model. You'd need to partition it across more than (350 GB) / (16 GB) ~ 22 GPUs just to run it! Training with the Adam optimizer (as they mention) would require at least 3 times as many (~66 GPUs), plus extra space for the activations. There are more memory-efficient optimizers though.

But there are **8 models** in the paper, **4 of which are smaller than GPT-2**, so some of those will probably be useful if OpenAI chooses to release them. 🙂

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Table 2.1: Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

👍 8   😄 1   🎉 2   👀 1

Reproducibility ✅

3. Running local models

deepseek-r1
ollama run deepseek-r1

⬇ 44.2M Downloads   🕐 Updated 3 months ago

DeepSeek's first-generation of reasoning models with comparable performance to OpenAI-o1, including six dense models distilled from DeepSeek-R1 based on Llama and Qwen.

1.5b   7b   8b   14b   32b   70b   671b

7b   🏷 29 Tags

Updated 3 months ago   0a8c26691023 · 4.7GB

model   arch qwen2 · parameters 7.62B · quantization Q4_K_M   4.7GB

params   { "stop": ["<|begin_of_sentence|>", "<|end_of_sentence|>",   148B

template {{- if .System }}{{ .System }}{{ end }} {{- range $i, $_ := …   387B

license   MIT License Copyright (c) 2023 DeepSeek Permission is hereby …   1.1kB

Privacy?

When you use API-based models, **your data (text, image, video) is shared with the companies** providing these services.

Privacy?

When you use API-based models, **your data (text, image, video) is shared with the companies** providing these services.

Privacy: Extremely sensitive data

Parent-Child Interaction (PCI)
`Version 1.0`

Youth of Utrecht, 2024, "Parent-Child Interaction (PCI)", https://doi.org/10.60641/7tty-4e58, ODISSEI Portal, V1

Cite Dataset ▾   Learn about Data Citation Standards.

Contact Owner   Share

Dataset Metrics ❔
0 Downloads ❔

Description ❔   Parent child interaction (PCI) is an observation of a parent with their child, with the goal to code qualitative aspects of the observed interaction based on explicitly defined behaviors. The PCI consists of age appropriate structured tasks that include a common mildly stressful event (teaching tasks and clean-up), and a pleasant event (unstructured free play).

Subject ❔   Social Sciences

Keyword ❔   communication, compliance, coping, parent-child relationship, parent child interaction, parental authority, parental behavior, parenting, free play, social interaction, language, sensitive discipline, parental sensitive discipline

Privacy: Extremely sensitive data

Parent-Child Interaction (PCI)

Version 1.0

Youth of Utrecht, 2024, "Parent-Child Interaction (PCI)", https://doi.org/10.60641/7tty-4...

Cite Dataset ▾   Learn about Data Citation Standards.

Description ❔   Parent child interaction (PCI) is an observation of a parent with their c...
qualitative aspects of the observed interaction based on explicitly defi...
of age appropriate structured tasks that include a common mildly stres...
clean-up), and a pleasant event (unstructured free play).

Subject ❔   Social Sciences

Keyword ❔   communication, compliance, coping, parent-child relationship, parent ...
authority, parental behavior, parenting, free play, social interaction, lan...
parental sensitive discipline

Secure ANalysis Environment (SANE)

SANE is a closed-off virtual machine where data providers authorise researchers to analyse their sensitive data using pre-approved software like R and Python. This environment ensures complete control for data providers and secure analysis for researchers.

Working with sensitive data is challenging for both researchers and data providers. Researchers often struggle with complex access procedures, while data providers may lack the infrastructure or expertise to offer secure data access. As a result, there is no shared platform that makes it easy to find, share, and process sensitive data securely and efficiently.

Ollama

1. **Downloads** Modelfile: model weights, configuration, and data

2. Supports **quantized** models

3. Highly optimized backend (llama.cpp)
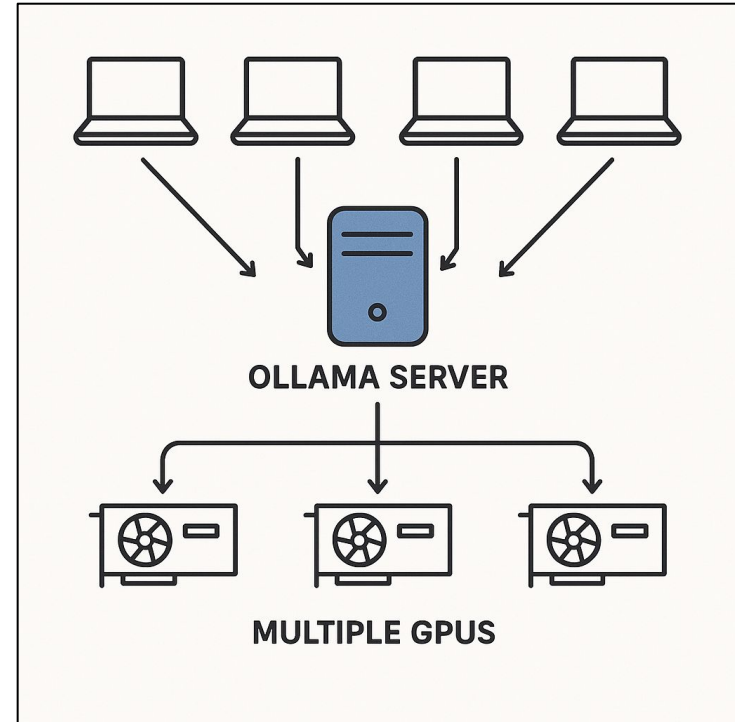

Get up and running with large language models.

Run Llama 3.3, DeepSeek-R1, Qwen 3, Mistral, Gemma 3, and other models, locally.

Download ↓

Available for macOS, Linux, and Windows

Ollama

4. Serve different models

5. Handle concurrent requests per model
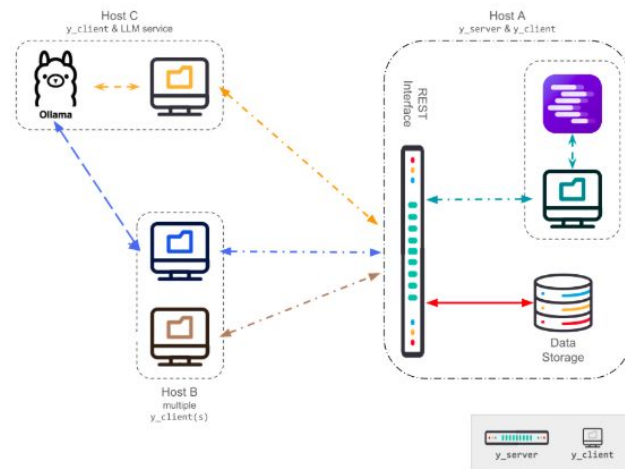
6. Fitting large models to multiple GPUs

Why Social Sciences Need HPC Infrastructure?

Example: Running <u>Llama (3.1 8B)</u> agent simulation of online communities

Resources: Paradox-V, 1 x NVIDIA A30

Simulation: 1000 users, 100 days

Runtime: <u>11 days</u>

CTRUST #7416 Prizma

Why Social Sciences Need HPC Infrastructure?

SPIN-Bench: How Well Do LLMs Plan Strategically and Reason Socially?

Model	Classical Planning		Competitive Games					Collaborative: Hanabi			
	Plan Acc ↑	N-Step↑	TTT$_{WR}$↓	C4$_{WR}$↓	CH$_{WR}$↓	C4$_{T3}$↑	CH$_{T3}$↑	2P↑	3P↑	4P↑	5P↑
o1	**58.59**	**16.09**	**30.0**	100.0	100.0	83.1	45.9	**16.4**	14.8	**14.8**	**14.2**
o1-mini	13.20	1.95	50.0	100.0	100.0	**87.0**	36.5	6.8	7.4	11.4	10.2
o3-mini	51.25	13.04	80.0	100.0	100.0	74.2	**52.8**	8.8	7.6	8.8	8.0
GPT-4o	8.75	0.60	100.0	100.0	100.0	84.1	32.2	6.6	4.8	4.8	4.6
GPT-4-turbo	5.62	0.13	40.0	100.0	100.0	83.8	38.7	5.2	5.6	5.0	6.0
Claude 3.5 Sonnet	20.55	4.44	40.0	100.0	100.0	78.9	49.5	8.2	9.4	7.4	8.4
Claude 3.5 haiku	4.22	0.30	50.0	100.0	100.0	69.6	35.9	2.4	4.0	2.8	2.8
DeepSeek R1	44.30	10.71	90.0	100.0	100.0	78.9	47.8	6.0	**16.0**	11.3	13.0
Llama-3.3-70b	5.78	0.32	100.0	100.0	100.0	79.5	25.4	N/A	N/A	N/A	N/A

Yao, J., Wang, K., Hsieh, R., Zhou, H., Zou, T., Cheng, Z., Wang, Z., & Viswanath, P. (2025).

SPIN-Bench: How well do LLMs plan strategically and reason socially? *[cs.AI]*.

arXiv. http://arxiv.org/abs/2503.12349

Why Social Sciences Need HPC Infrastructure?

Example: Running <u>Deepseek R1</u> agent simulation of online communities

Simulation: 1M users, 1 year

Resources: Paradox-V+++, 15 x NVIDIA H100

Why Social Sciences Need HPC Infrastructure?

Ursula von der Leyen
President of the European Commission

Dear Professor,

We have stumbled upon €2 billion of unallocated funds in the EU budget. We would like to grant these funds to your institution. Could you please reply with a quick budget proposal outlining how you would use these funds?

Warm regards,
Ursula von der Leyen
President of the European Commission

1. Localized AI for **trustworthy** and advanced CSS

2. CSS **participation** in the development on joint
   HPC & Localized AI Infrastructure

Thank you!

atomashevic@ff.uns.ac.rs

www.atomasevic.com