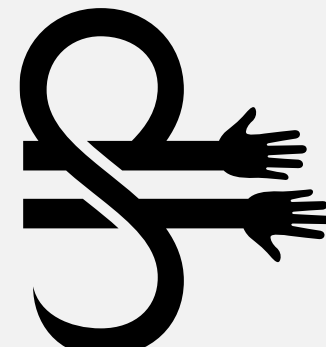


# PRIMJENA VELIKIH JEZICNIH MODELA U DRUŠTVENIM ZNANOSTIMA

Od oblikovanja uputa do  
pouzdanog i valjanog  
mjernog procesa

Izrađeno prema materijalima radionice Giovannija Pagana  
„Large Language Models for Social Science Research”,  
održane u sklopu ljetne metodološke škole Europskog  
konzorcija za politička istraživanja (ECPR), rujan 2026.



demohubu

# SADRŽAJ

1.  
Mogućnosti LLM-ova  
u društvenim  
znanostima

2.  
Logika LLM-ova

3.  
Vrste modela i  
infrastruktura

4.  
Dizajn radnog tijeka i  
prompt osjetljivost

5.  
Validacija i  
reproducibilnost

6.  
Praktične smjernice i  
primjeri

# MOGUĆNOSTI LLM-OVA U DRUŠTVENIM ZNANOSTIMA

# VAŽNOST TEKSTA

- Društveni i politički život ostavljaju obimne tekstualne tragove
- Govori, manifesti, intervjui, sudske odluke, *policy* dokumenti, rezultati anketa, društveni mediji...
- Stari problem: tekst je izdašan, ali pažljivo čitanje oskudno
- LLM-ovi ne zamjenjuju čitanje (ili barem ne bi trebali) oni ga pojačavaju sustavnim mjerenjem

# MOGUĆNOSTI PRIMJENE LLM-OVA

- Sposobnost predviđanja sljedećeg tokena može se iskoristiti na mnoge načine, pod uvjetom da se zadatak može svesti na tekst
- Analiza teksta: Klasifikacija, ekstrakcija, skaliranje
- Kvalitativno kodiranje: Pomoć pri sistematskom kodiranju
- Pronalaženje: Pronalaženje relevantnih dijelova
- Simulacija: Agentni modeli društvenih dinamika
- Ankete: Predviđanje, sintetski odgovori
- Eksperimenti: Generiranje stimulusa, simulacija

# ČETIRI PRINCIPA AUTOMATSKE ANALIZE TEKSTA

*(Grimmer i Stewart, 2013)*

1.

- Svi kvantitativni modeli jezika su pogrešni, ali neki su korisni

2.

- Kvantitativne metode za tekst dopunjuju i pojačavaju ljudske sposobnosti, ali ih ne zamjenjuju

3.

- Ne postoji globalno najbolja metoda za automatsku analizu teksta

4.

- Validacija, validacija, validacija

# 1. SVI KVANTITATIVNI MODELI JEZIKA SU POGREŠNI – ALI NEKI SU KORISNI

- Kvantitativni modeli ne odražavaju stvarni proces generiranja jezika
- Model je svrhovito pojednostavnjenje
- Zadržava vidljivima određena obilježja, a druga potiskuje u pozadinu
- Modeli funkcioniraju na isti način, ali na drugoj razini

## 2. KVANTITATIVNE METODE ZA TEKST DOPUNJUJU I POJAČAVAJU LJUDSKE SPOSOBNOSTI, ALI IH NE ZAMJENJUJU

- Automatske metode ne eliminiraju potrebu za pažljivim i dubinskim čitanjem teksta, već pojačavaju analitičke kapacitete istraživača
- Ljudska prosudba ostaje prisutna i (ključna) u tijeku rada:
  - Izrada kodne knjige
  - Određivanje granica korpusa
  - Pregled uzoraka
  - Analiza pogrešaka
  - Donošenje odluka u složenim slučajevima
  - Izvještavanje o razini nesigurnosti

### 3. NE POSTOJI GLOBALNO NAJBOLJA METODA ZA AUTOMATSKU ANALIZU TEKSTA

- LLM sam po sebi nije metoda, a kamoli „najbolja”
- Izbor metode izravno ovisi o istraživačkom pitanju, vrsti podataka i ciljanoj mjeri
  - Klasificiranje
  - Sažimanje
  - Ekstrakcija
  - Prevođenje
  - Dohvaćanje
  - Generiranje
  - Simulacija

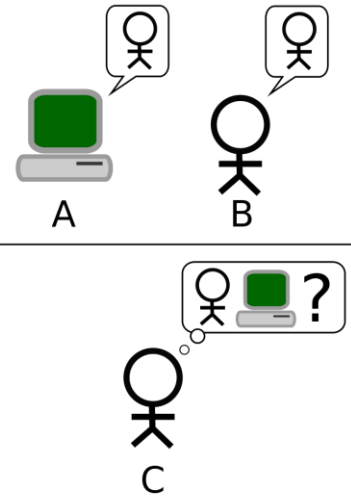
## 4. VALIDACIJA

- Sastavni dio SVAKOG nacрта istraživanja
- Budući da su modeli nesavršeni, nužna je sustavna validacija rezultata u odnosu na konceptualne ciljeve i ljudske ocjene
- Što je izlazni rezultat modela važniji za samu tvrdnju, to je potrebno više dokaza za njegovu validaciju

# LOGIKA LLM-OVA

# LLM LOGIKA

- Modeli su veliki s obzirom na **količinu memorije** potrebnu za njihovu upotrebu
- Sklopovi koji na temelju niza ulaznih riječi pogađaju izlaznu riječ
  - **Generative**: model je sposoban generirati nastavak na zadani unos; na temelju određenog teksta model pokušava predvidjeti koje riječi slijede
  - **Pre-trained**: model je obučen na vrlo velikom korpusu općeg teksta; zamisao je da se obuči jednom te koristi za brojne različite namjene, bez potrebe za ponovnim obučavanjem od nule (*fine-tuning*)
  - **Transformer**: specifična vrsta modela dubokog učenja arhitekture enkoder-dekoder, koji se temelji na samonadziranom učenju
- Razumijevanje semantičkih veza



# LLM JE STROJ ZA VJEROJATNOST, NE TRAŽILICA



- Cilj treniranja generativnih modela jest predviđanje uvjerljivih nastavaka (distribucija vjerojatnosti nad sljedećim tokenima)
- Metodološki oprez: ono što je vjerojatno nije nužno i točno

Istraživačko pitanje

Konceptualizacija

Operacionalizacija

Mjerenje

Validacija

# VRSTE MODELA I INFRASTRUKTURA

# KAKO KORISITI AI MODELE U ISTRAŽIVANJIMA

*Dvije vrste AI modela*

## Open-Weight

- Kreator objavljuje parametre istrenirane neuronske mreže („težine”), omogućujući preuzimanje samog modela
- Veća mogućnost podešavanja
- Inicijalni infrastrukturni troškovi, dugoročno isplativije
- Potpuna kontrola podataka
- Reproducibilnost
- Primjeri: DeepSeek, Llama, Mixtral...

## Closed API

- Kreator zadržava model za sebe i pokreće ga na vlastitoj infrastrukturi, nudeći pristup isključivo putem softverskog API-ja
- Manja mogućnost podešavanja
- Plaćanje pristupa, ne vlasništva (ovisnost o promjenama uvjeta i premija)
- Plaćanje bazirano na korištenju (prompt ili token)
- Podaci prolaze kroz vanjski server
- Primjeri: GPT, Claude, Gemini...

# OPEN SOURCE LLM-OVI

*Popis nije iscrpan niti predstavlja rang listu prema učinkovitosti*

1. Deepseek V3, R1, V3.1 and V3.2
2. GPT-OSS 20B, 120B (OpenAI)
3. Gemma initial release, 1.1, 2, 3, 3n (Googla)
4. Qwen initial relase, 2, 2.5, 3 (Qwen AI of Alibaba)
5. Mistral Small, Large 3 (Mistral)
6. Llama 1, 2, 3, 3.1, 3.2, 4 (Meta AI)
7. Kimi K1.5, VL, Dev, K2, Linear (Moonshot AI)
8. GLM 4.1, 4.5, 4.6 , 4.7 (Z AI)
9. Nemotron 2, 3 (NVIDIA)
10. Mimo 7B, V2-Flash (Xiaomi Mimo)

Gdje nabaviti modele?

# HUGGING FACE



<https://huggingface.co/>

- Platforma i biblioteka za primjenu modela strojnog učenja u širokom spektru zadataka, uključujući konverzacijsku umjetnu inteligenciju, obradu prirodnog jezika (NLP), računalni vid te analizu i klasifikaciju sentimenta



Za pokretanje većine LLM-ova dostupnih na platformi Hugging Face potreban je snažan grafički procesor (GPU)

- U suprotnom, na raspolaganju su opcije poput *cloud* platformi ili Google Colaba

Gdje pokrenuti modele?



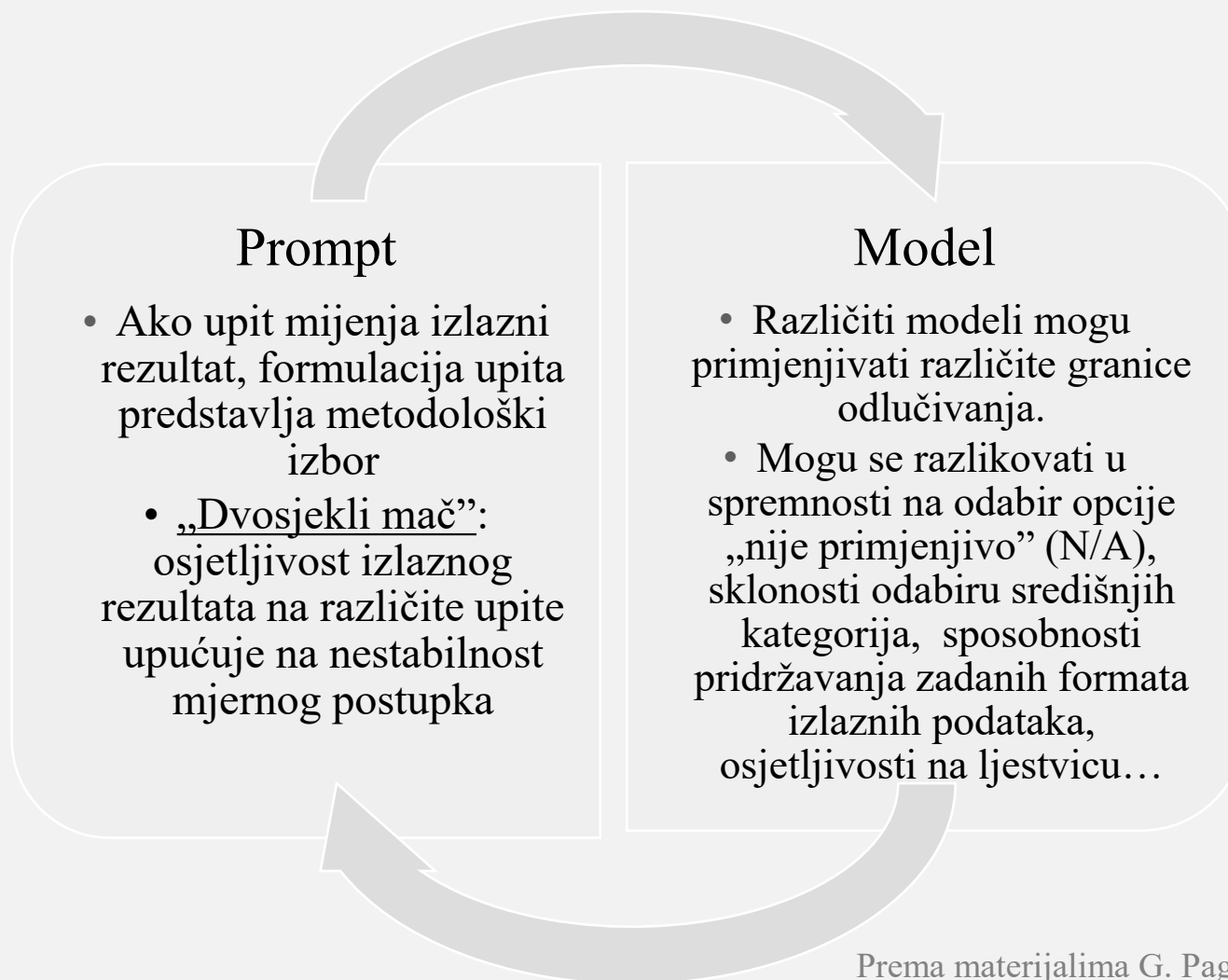
<https://colab.research.google.com/>

- Omogućuje pisanje i izvršavanje Python koda u pregledniku, uz:
  - Odsutnost potrebe za konfiguracijom
  - Besplatan pristup GPU-ovima
  - Jednostavno dijeljenje
- Osnovna logika
  - Učitavanje većeg otvorenog modela u Colabu
  - Definiranje zadatka
  - Prosljeđivanje izvornog teksta modelu
  - Pretvaranje odgovora u strukturirani objekt

# DIZAJN RADNOG TIJEKA I PROMPT OSJETLJIVOST

# PROMPT I MODEL KAO DIO METODE

*Različiti upiti i modeli utječu na izlazne rezultate*



## Iluzija jednostavnog mjerenja

- Opasnost leži u tome što uvjerljiv rezultat može prikriti manjkav proces mjerenja

# PISANJE PROMPTA

*Nekoliko (jako) generalnih savjeta i uputa za usmjeravanje LLM-a*

## 1. Specificiranje LLM-ove uloge

- Npr. kodiranje, sažimanje, ekstrakcija, generiranje, dohvaćanje, simulacija

## 2. Sužavanje i konkretiziranje tvrdnje na materijal koji istražujemo

- Usmjeravanje modela da „odgovara” samo na temelju teksta koji učitavamo

## 3. Definiranje onoga što se smatra dokazom

- Upute modelu vezane uz to kada se slučaj može kodirati, kako da tumači mjerni instrument (skalu, kodnu shemu i sl.), kada je dopuštena oznaka „nije primjenjivo” (N/A)
- Ne postoji najbolji prompt: odabir upita i modela u interakciji je sa zadatkom

**Predmemoriranje/ zapisivanje (logging):**  
pohranjivanje upita, izvornih izlaza, obrađenih izlaza, naziva modela i datuma (provjera pravila o postupanju)

Istraživačko pitanje

Konceptualizacija

Operacionalizacija

Mjerenje

Validacija

# OPREZ VEZAN UZ N/A

*Ako je riječ o velikom obujmu analiziranog teksta, model bi trebao „izbaciti” nekoliko N/A ocjena*

1. Oznaka N/A trebala bi značiti da analizirani tekst možda ne sadrži dovoljno dokaza o predmetnom pitanju
2. **Ako oznake N/A potpuno izostaju, to je „red flag”**
3. To što model nikada ne vraća N/A ne znači nužno da je svaku jedinicu analiziranog teksta moguće kodirati – to može značiti da model daje odgovor čak i uz vrlo oskudne dokaze ili njihov potpuni izostanak (halucinacije)
4. Ako nema N/A oznaka, nije loše provjeriti i moguću pojavu „inflacije središnje vrijednosti”, u tom je slučaju model ono što je trebalo biti N/A ocijenio kao srednju vrijednost

When you did see a few red flags but you're sure you can change him:



# KAKO (PROBATI) OSIGURATI DOBAR OUTPUT

1. Provođenje pilot-projekta prije prelaska na masovnu obradu
  - a. Jedan prompt s jednom jedinicom analize
  - b. Isti prompt s drugom jedinicom
  - c. Ista jedinica s drugim promptom
  - d. Obrada manjeg udjela ukupnog uzorka i usporedba s ljudskom obradom
2. Analiza stope (ne)uspjeha raščlambe (parsinga)
3. Usporedba barem dva prikladna modela
4. Spremanje podataka u predmemoriju (cache) nakon svakog retka
5. Dizajniranje skripte tako da može nastaviti s radom nakon prekida

Istraživačko pitanje

Konceptualizacija

Operacionalizacija

Mjerenje

Validacija

# VALIDACIJA I REPRODUCIBILNOST

# POUZDANOST I VALJANOST

*Odgovaraju na različita pitanja*

## Dokazi za pouzdanost

- Dva ljudska koda dodjeljuju iste oznake
- Model dodjeljuje istu oznaku pri ponovnom pokretanju s determinističkim postavkama
- Rezultati su stabilni unatoč manjim izmjenama u formulaciji upita
- Kodiranje je slično na različitim uzorcima dokumenata
- Stupanj slaganja između ljudskih koda je visok
- Stupanj slaganja između modela i ljudi je visok

Daje li postupak mjerenja dosljedne rezultate?

## Dokazi za valjanost

- Valjanost sadržaja: utvrđivanje da kodna knjiga pokriva namjeravani konstrukt
- Konvergentna valjanost: koreliranost s drugom utvrđenom mjerom istog konstrukta
- Diskriminantna valjanost: ne postojanje snažne korelacije s teoretski različitim konstruktima
- Valjanost poznatih skupina: slučajevi za koje se teoretski zna da se razlikuju na konstrukt dobivaju različite rezultate
- Valjanost prediktivnog kriterija: mjera predviđa ili se odnosi na vanjski kriterij koji bi teoretski trebala predvidjeti

...

Znače li dobivena mjerenja ono što tvrdimo da znače?

# VALIDACIJSKI PROTOKOL

*Što bi metodološki dio rada trebao sadržavati i na koja si pitanja moramo odgovoriti*

- Svrha istraživanja: što ćete na temelju rezultata moći ustvrditi ili odlučiti?
- Izvor i jedinica: koji se materijal obrađuje i što se smatra jednim slučajem?
- Pravilo za output: oznaka, bodovna vrijednost, sažetak, polje za ekstrakciju, kod, bilješka...
- Provjera: kako se verificira manji broj slučajeva (cca 10%)?
- Slučajevi neuspjeha: koje bi pogreške rezultat učinile neuporabljivim?
- Rečenica za izvještavanje: što biste rekli skeptičnom recenzentu?

# THE GUIDE-LLM CHECKLIST FOR PROMOTING TRANSPARENCY, REPRODUCIBILITY AND ETHICAL ACCOUNTABILITY IN STUDIES THAT INVOLVE LLMS

<https://www.nature.com/articles/s41562-026-02492-7>

| Item                                 | Name                                                                                                                                |
|--------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------|
| <b>Scope of LLM use</b>              |                                                                                                                                     |
| Item A.1                             | LLMs were used in this project for: ...                                                                                             |
| Item A.2                             | Degree of automation (human in the loop versus fully automated): ...                                                                |
| <b>Model or system details</b>       |                                                                                                                                     |
| Item B.1                             | Model name, including provider, model size, exact version or identifier, date of access and source link (if possible): ...          |
| Item B.2                             | Model access (for example, API, web interface or local) and context mode (for example, chat mode or separate calls): ...            |
| Item B.3                             | Relevant LLM configuration(s) reported (as applicable), such as temperature, maximum tokens, seed and number of runs: ...           |
| Item B.4                             | Customization: ...                                                                                                                  |
| Item B.5                             | Did the LLM session(s) include persistent memory across interactions? ...                                                           |
| <b>Prompts</b>                       |                                                                                                                                     |
| Item C.1                             | Exact prompt(s) reported: ...                                                                                                       |
| Item C.2                             | System-wide instructions (if any): ...                                                                                              |
| <b>Data inputs and privacy</b>       |                                                                                                                                     |
| Item D.1                             | Handling of personal or sensitive data (if any) (for example, consent for data processing): ...                                     |
| <b>Validation and interpretation</b> |                                                                                                                                     |
| Item E.1                             | Human validation of LLM outputs: ...                                                                                                |
| Item E.2                             | Describe any relevant post-processing (for example, filtering in case of format mismatches, unit conversions and so on): ...        |
| <b>Reproducibility</b>               |                                                                                                                                     |
| Item F.1                             | Code, notebooks or scripts for LLM calls shared: ...                                                                                |
| <b>Competing interests</b>           |                                                                                                                                     |
| Item G.1                             | Funding, support or other relevant relationships (including in-kind access to compute or models, or professional affiliations): ... |

# PRAKTIČNE SMJERNICE I PRIMJERI

# UPUTE ZA RAD U LLM-OVIMA

- Složeni se materijal rijetko mjeri u jednom koraku
- Potreban je pripremni rad, koji rezultira:
  - Segmentima
  - Dimenzijama
  - Kodovima
  - Primjerima
  - Graničnim slučajevima
  - Pilot-oznakama

# DEMOHUB PRIMJER

*Bilješka o protokolu istraživanja uz pomoć LLM-a*

| Stavka                  | Opis                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
|-------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Research question/claim | How is the concept of "democratic school culture" conceptualized and operationalized across international educational research, and what key dimensions (e.g., governance, curriculum, student agency, equity) dominate the literature?                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| Role of the LLM/task    | The LLM serves as an intermediate extraction and classification tool upstream of the synthesis. It parses article segments to identify explicit conceptual definitions, extract verbatim supporting quotes, and categorize them into primary theoretical dimensions to reduce manual filtering time across a large corpus.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| Data                    | <i>Source:</i> 198 core peer-reviewed journal articles selected from an initial pool of 26,224 WoS/Scopus records<br><i>Size:</i> ~3,000–5,000 individual paragraphs/text segments extracted from full-text PDFs<br><i>Unit of Analysis:</i> Paragraphs containing definition claims, conceptual frameworks, or operationalizations<br><i>Languages &amp; Countries:</i> English; international literature with a focus on European educational contexts<br><i>Time Period:</i> 1960–2025                                                                                                                                                                                                                                                                                                                        |
| Prompt                  | <p>Plaintext</p> <p>You are an expert qualitative research assistant in educational sociology.</p> <p>Analyze the following text segment from a peer-reviewed article.</p> <p>Determine if the text explicitly defines or operationalizes "democratic school culture" or "democratic education".</p> <p>Return ONLY a valid JSON object matching this schema:</p> <pre>{ "has_definition": true/false,<br/>  "verbatim_quote": "Exact quote from text containing the definition, or null if false",<br/>  "conceptual_dimension": "Governance"   "Curriculum"   "Student Agency"   "Equity/Access"   "Unclear",<br/>  "reasoning": "1-2 sentence explanation for the selected dimension",<br/>  "confidence": "High"   "Medium"   "Low" }</pre> <p>Text Segment:</p> <p>""""[INSERT PARAGRAPH TEXT HERE]""""</p> |

# DEMOHUB PRIMJER

*Bilješka o protokolu istraživanja uz pomoć LLM-a*

| Stavka                | Opis                                                                                                                                                                                                                                                                                                                                                                                                         |
|-----------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Validation            | Constructing a reference benchmark of 30 randomly sampled segments independently coded by two human researchers using a codebook. Inter-coder reliability (Cohen's kappa) will be established. Model performance will be evaluated against this adjudicated ground truth using precision, recall, and macro-F1 across dimensions, along with qualitative error analysis for hallucinations or rule breaches. |
| Baseline              | Comparison against a traditional keyword/dictionary search (e.g., string matching for "democrat*", "governance", "culture") to demonstrate that the LLM improves precision and handles context far better than rule-based pattern matching.                                                                                                                                                                  |
| Risks                 | Conceptual conflation (e.g., treating general "school climate" or "positive environment" as "democratic culture"), hallucinating or altering verbatim quotes, and collapsing overlapping dimensions.                                                                                                                                                                                                         |
| Reproducibility       | Shared repository containing Python/API scripts, exact prompt templates, model checkpoint identifier (e.g., gpt-4o-2024-08-06 or Ministral-3-14B-Instruct), decoding parameters (temperature = 0.0, seed), JSON schema definitions, and the anonymized validation dataset.                                                                                                                                   |
| Ethics                | All source material consists of published, public academic literature; no private human subject data, restricted survey data, or personally identifiable information (PII) is involved.                                                                                                                                                                                                                      |
| Expected contribution | Provides a scalable, systematic mapping of how democratic school culture is defined globally, identifying under-researched dimensions and establishing a transparent methodological protocol for LLM-assisted systematic literature reviews in education.                                                                                                                                                    |