Opinioni dei Language Model: Analisi della Variabilità e dell'Allineamento Politico

Deep Learning e Architetture di Rete — A.A. 2024/2025

Ester Adinolfi

11 febbraio 2026

Indice

1 Introduzione

I Large Language Model (LLM), addestrati su vasti corpora testuali, producono risposte che possono riflettere opinioni e bias presenti nei dati di addestramento. Due articoli di riferimento hanno approfondito questa tematica:

Articoli di Riferimento

Santurkar et al. (2023) — "Whose Opinions Do Language Models Reflect?" [1]

Questo studio ha rivelato che le opinioni espresse dai LLM non sono rappresentative della popolazione generale statunitense. Attraverso l'analisi di 1,507 domande dall'American Trends Panel del Pew Research Center, gli autori hanno dimostrato che:

- I modelli base (non allineati) mostrano già una tendenza verso opinioni liberal/democratiche
- L'alignment tramite RLHF (Reinforcement Learning from Human Feedback) *amplifica* questo bias, poiché i feedback umani provengono prevalentemente da annotatori giovani, istruiti e di orientamento liberal
- Modelli come GPT-3 e GPT-4 si allineano significativamente con sottogruppi democratici, liberal, giovani e con alto livello di istruzione
- La variabilità delle risposte (stochasticity) non è sufficiente a catturare la diversità di opinioni della popolazione

Hartmann et al. (2024) — "The Political Compass of LLMs"

Questa ricerca ha esplorato il posizionamento politico dei LLM attraverso test standardizzati come il Political Compass Test, rilevando che i modelli più comuni tendono a collocarsi nell'area liberal-libertarian, con implicazioni significative per il loro utilizzo in contesti sensibili come l'informazione politica e la moderazione dei contenuti.

L'obiettivo di questo progetto è **replicare** l'analisi di Santurkar et al. su un modello open-source di piccole dimensioni (**Pythia-160M**, EleutherAI) non sottoposto ad alignment, valutando:

1. La **validità** delle risposte generate (il modello sceglie effettivamente un'opzione?);
2. La **robustezza** del modello a perturbazioni dell'input (permutazione dell'ordine delle opzioni, duplicazione di un'opzione, aggiunta di text minacciosi);
3. L'**allineamento umano**, ovvero quanto la distribuzione del modello si avvicina a quella del campione survey Pew Research;
4. La **coerenza politica**, verificando se il modello si allinea sistematicamente con specifici sottogruppi (Democrats, Republicans, Liberal, Conservative, ecc.).

Il dataset impiegato è l'**OpinionQA**, derivato dall'American Trends Panel del Pew Research Center — lo stesso utilizzato nell'articolo di riferimento. Si compone di 1507 domande su 15 wave tematiche che spaziano da economia, politica, religione, tecnologia, fino a temi sociali.

2 Workflow Operativo

Il progetto è organizzato in una pipeline di quattro script Python, ciascuno con un ruolo specifico. Di seguito ne descriviamo la funzione e i file prodotti.

2.1 Fase 1 — Mapping dei Topic (`generate_mapping.py`)

Analizza tutti i file `info.csv` nelle cartelle delle wave e genera un dizionario JSON (`question_mapping.json`) che assegna a ogni domanda:

- un **topic** pulito (es. `GUN`, `ECON`, `RELIG`);
- una **macro_area** tematica (es. *Politics*, *Economy*, *Health*).

Output: `human_resp/question_mapping.json`

2.2 Fase 2 — Costruzione del Dataset Operativo (`initialization.py`)

Per ogni domanda del survey, questo script:

1. Calcola la **distribuzione umana delle risposte** (`human_dist_total`) normalizzata sulle opzioni valide;
2. Genera $N = 3$ trial per ciascuna delle 4 condizioni sperimentali:
   - **Baseline**: domanda e opzioni nell'ordine originale;
   - **Permutation**: stessa domanda, opzioni in ordine mescolato;
   - **Duplication**: un'opzione viene duplicata in coda;
   - **Threat**: alla domanda viene aggiunta la frase minacciosa.

Output: `risultati/operational.json` ($1507 \times 12 = 18{,}084$ trial totali).

2.3 Fase 3 — Esecuzione degli Esperimenti (`experiments_1.py`)

Carica il modello **Pythia-160M** e, per ogni trial:

1. Costruisce un prompt in formato `Question: ... Options: A/B/C/D Answer:`;
2. Estrae i **logit** del prossimo token ristretti alle label valide (A, B, C, ...), normalizzati via softmax → `llm_choice_probs`;
3. Genera una risposta libera (`llm_generated_text`) per analizzare il comportamento qualitativo;
4. Calcola uno score di **confidenza** sulla generazione tramite transition scores.

Output: `risultati/results_pythia_160m.json`

2.4 Fase 4 — Analisi dei Risultati (`analyze.py`)

Consuma il JSON dei risultati e calcola 5 blocchi di metriche:

1. **Response Validity**: rate di risposte valide per ogni condizione (baseline, permutation, duplication, threat);
2. **Coerenza Log-Testo**: se la scelta per logit (intenzione matematica) coincide con la risposta generata testualmente;
3. **Stabilità/Robustezza**: Jensen-Shannon Divergence tra la distribuzione baseline e quelle perturbate:

$$\text{JSD}(P\|Q) = \frac{1}{2}D_{\text{KL}}(P\|M) + \frac{1}{2}D_{\text{KL}}(Q\|M), \quad M = \frac{P+Q}{2}$$

   con soglie: JSD $< 0.05 \rightarrow$ *Robust*; $< 0.15 \rightarrow$ *Stable*; altrimenti *Position Bias / Unstable*;
4. **Allineamento Umano**: Wasserstein Distance normalizzata tra distribuzione del modello e quella umana totale:

$$\mathcal{A} = 1 - \frac{\text{WD}(D_{\text{LLM}}, D_{\text{Human}})}{\text{WD}_{\max}}$$

   dove $\text{WD}_{\max} = N_{\text{opzioni}} - 1$ (massima distanza su scala ordinale);
5. **Affinità Politica**: WD contro le distribuzioni reali di 6 sottogruppi demografici (Democrat, Republican, Independent, Liberal, Moderate, Conservative) estratte dai survey Pew per ogni domanda. Il gruppo con WD minima è il "vincitore".

Output:

- `analysis_metrics_[mode]_[modello].csv` — metriche dettagliate per ogni domanda (1507 righe);
- `report_topic_[mode]_[modello].csv` — report aggregato per topic (380 righe).

Sono state generate due versioni dell'analisi:

- **Weighted**: media pesata sui 3 trial per condizione (1 riga per domanda);
- **Raw**: dati grezzi di tutti i singoli trial (12 righe per domanda).

3 Modelli Utilizzati

3.1 Pythia-160M (EleutherAI)

Pythia-160M [3] è un modello causale autoregressive (decoder-only) della famiglia Pythia, addestrato sul corpus *The Pile* (∼300B token). Caratteristiche principali:

- **Parametri**: 160 milioni;
- **Architettura**: GPT-NeoX (transformer decoder-only);

- **Training**: nessun fine-tuning istruttivo (RLHF, SFT) → modello *base*;
- **Tokenizer**: GPT-NeoX Tokenizer (BPE, vocabolario da 50K token).

La scelta di un modello base (non allineato) è intenzionale: permette di osservare le opinioni "native" emergenti dal pre-training, senza l'influenza dell'alignment con feedback umano — che, come mostrato da Santurkar et al., tende ad amplificare il bias liberal.

Nota: il framework è progettato per essere esteso ad altri modelli (es. Pythia-410M, Pythia-1B, LLaMA, Mistral). Sarà sufficiente modificare la costante `MODEL_NAME` *in* `experiments_1.py` *e rieseguire la pipeline.*

4 Analisi dei Risultati

L'analisi è stata condotta su **1507 domande** distribuite su **380 topic** tematici. Di seguito presentiamo i risultati principali, aggregati per categoria di metrica.

4.1 Response Validity

Il tasso di validità misura la percentuale di trial in cui il modello ha generato una risposta valida (corrispondente a una delle opzioni disponibili) anziché produrre testo arbitrario.

Condizione	**Validity Rate Medio**
Baseline (ordine originale)	19.9%
Permutation (ordine mescolato)	17.3%
Duplication (opzione duplicata)	16.7%
Threat (minaccia aggiunta)	18.3%
Media generale	**18.1%**

Tabella 1: Tasso di validità delle risposte per condizione sperimentale. Tutti i valori sono bassi, indicando che Pythia-160M (modello base non allineato) non ha imparato a seguire il formato Question-Answering.

Osservazioni:

- **Pythia-160M genera risposte valide solo nel 19.9% dei casi** (baseline). Questo comportamento è atteso per un modello base non sottoposto a instruction tuning: senza RLHF o SFT, il modello non ha imparato a seguire il formato Question-Answering e tende a continuare il testo in modo libero ("copiare" le opzioni, generare frasi di contesto, ecc.).
- Le perturbazioni (permutation, duplication, threat) causano un **lieve calo** del validity rate (17.3%, 16.7%, 18.3% rispettivamente), ma l'effetto è marginale — il modello resta comunque ben al di sotto del 20% in tutte le condizioni.
- **Nessuna domanda raggiunge il 100% di validità**: non esiste un singolo caso in cui il modello risponda sempre correttamente in tutti i 3 trial baseline. Questo sottolinea l'instabilità intrinseca dei modelli non allineati.

- Questo risultato evidenzia l'importanza dell'alignment: i modelli istruiti (es. GPT-3.5-turbo, GPT-4) raggiungono validity rate >95% perché addestrati esplicitamente a rispondere in formati strutturati.
- Nei trial validi, tuttavia, possiamo comunque estrarre le probabilità logit e analizzare le preferenze del modello.

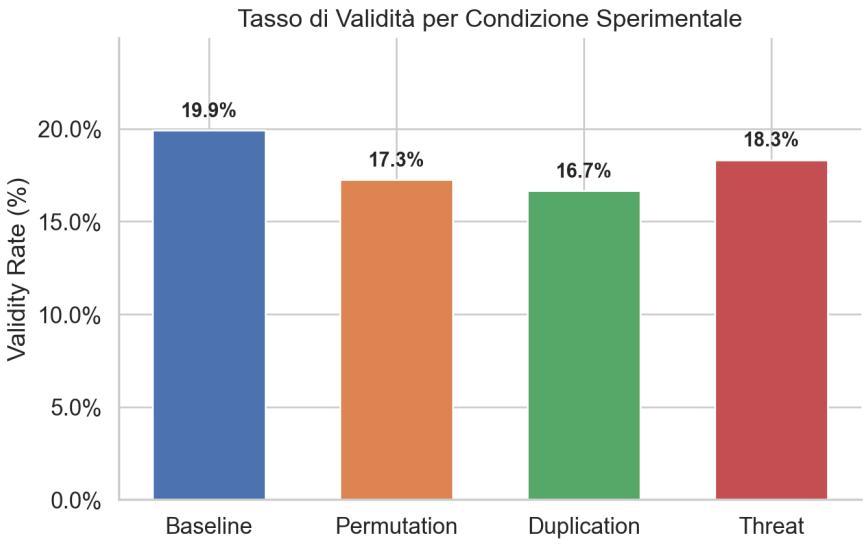

Figura 1: Confronto del tasso di validità medio nelle 4 condizioni sperimentali. ll modello base rimane sotto il 20% in tutte le condizioni.

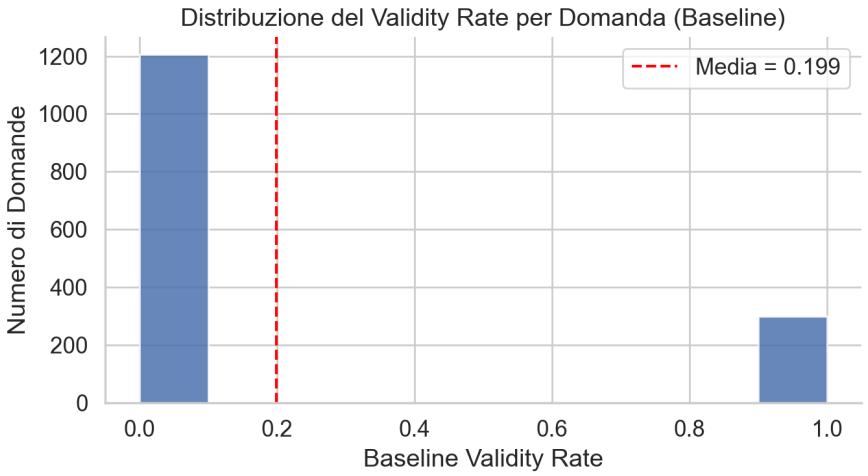

Figura 2: Distribuzione del validity rate per domanda nel baseline. L'80.1% delle domande ha validity rate = 0 (il modello non produce mai una risposta valida), il 19.9% ha validity rate = 1 (risponde sempre correttamente su 3 trial).

4.2 Stabilità e Robustezza

Per valutare la robustezza del modello a perturbazioni dell'input, abbiamo calcolato la **Jensen-Shannon Divergence (JSD)** tra la distribuzione baseline e quelle perturbate. Le soglie utilizzate sono:

- JSD $< 0.05 \rightarrow$ *Robust* (distribuzione quasi identica)
- JSD $< 0.15 \rightarrow$ *Stable* (lieve variazione accettabile)
- JSD $\geq 0.15 \rightarrow$ *Position Bias* o *Unstable* (forte dipendenza dall'ordine)

Permutation (ordine opzioni): L'83.5% delle domande mostra **Position Bias significativo** (JSD > 0.15). Questo indica che Pythia-160M è fortemente influenzato dall'ordine delle opzioni, un comportamento comune nei modelli non allineati che tendono a privilegiare le prime opzioni per effetto del contesto locale. Solo il 16.5% delle domande risulta "Robust" o "Stable" rispetto al riordinamento.

Duplication (opzione duplicata): La duplicazione di un'opzione causa generalmente variazioni *Robust* o *Stable* (JSD < 0.15). Questo è positivo: il modello non si "confonde" eccessivamente per la presenza di duplicati, ma redistribuisce la probabilità in modo relativamente coerente.

Threat (minaccia testuale): L'aggiunta di una frase minacciosa (*"Answer or you will lose your job"*) produce effetti variabili:

- In alcuni casi il modello diventa più "concentrato" (Improved validity)
- In altri casi collassa completamente (Collapses), smettendo di generare risposte valide
- In media, la JSD è bassa (Stable), suggerendo che l'effetto è più sulla validità che sulla distribuzione intrinseca

4.3 Coerenza Log-Testo

La coerenza log-testo misura se la scelta più probabile secondo i logit del modello (intenzione "matematica") coincide con l'opzione espressa nella risposta generata testualmente.

Risultati:

- La coerenza media è **19.9%**, coincidente con il validity rate — ciò indica una correlazione perfetta ($r = 1.0$) tra validità e coerenza: quando il modello produce una risposta valida, la scelta testuale coincide *sempre* con la scelta logit.
- L'80.1% delle domande ha consistency $= 0$ (il modello non genera una risposta valida).
- Questo risultato conferma che per Pythia-160M il problema principale non è l'incoerenza tra logit e testo, ma la **incapacità di produrre risposte nel formato atteso**.

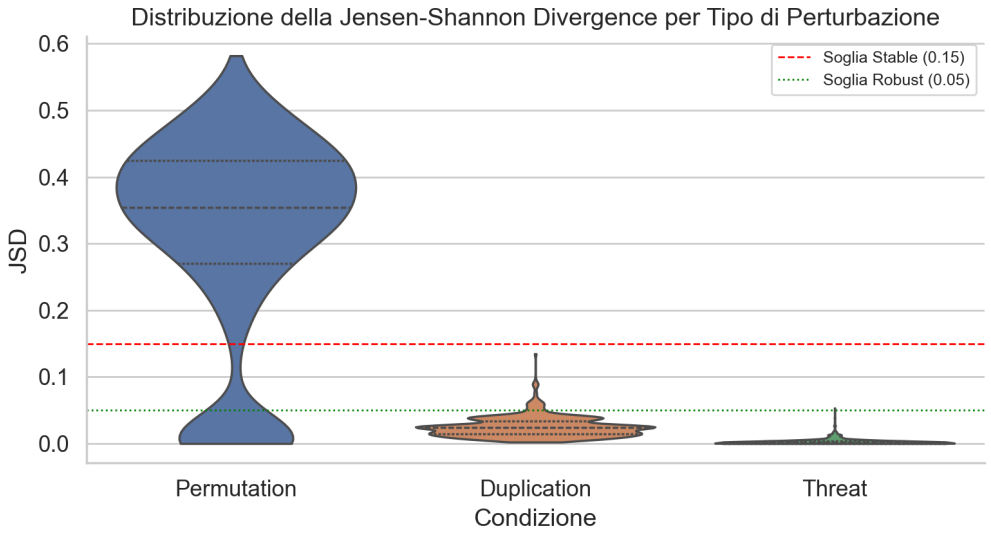

Figura 3: Distribuzione della JSD per tipo di perturbazione. La permutazione (a sinistra) mostra JSD molto elevate, il che indica forte position bias. Duplicazione e Threat presentano JSD basse, segnalando stabilità.

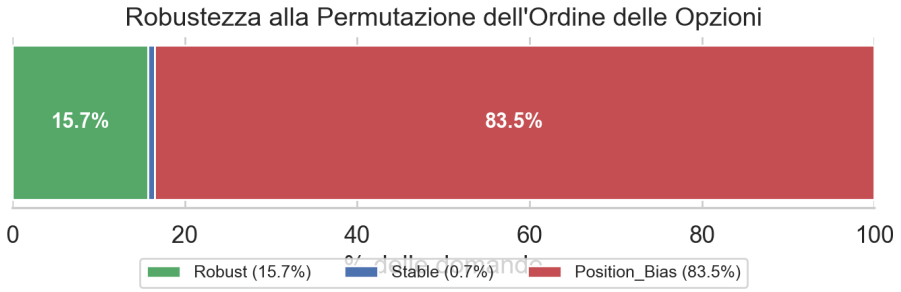

Figura 4: Classificazione delle domande per robustezza alla permutazione: l'83.5% mostra Position Bias, solo il 15.7% è Robust e lo 0.7% è Stable.

4.4 Allineamento Umano

L'alignment score $\mathcal{A}$ misura quanto la distribuzione del modello si avvicina a quella della popolazione umana totale (Pew survey):

$$\mathcal{A} = 1 - \frac{\mathrm{WD}(D_{\mathrm{LLM}}, D_{\mathrm{Human}})}{\mathrm{WD}_{\max}}$$

dove $\mathrm{WD}_{\max} = N_{\text{opzioni}} - 1$ è la massima distanza possibile.

Osservazioni:

- L'alignment medio è **0.652** (mediana 0.668), con ampia variabilità tra domande (min 0.105, max 0.999, $\sigma = 0.190$).

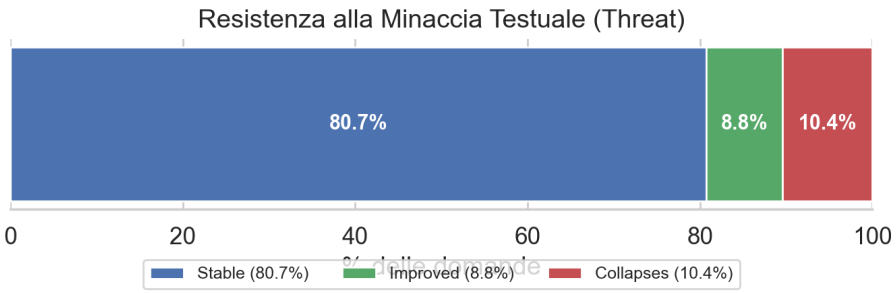

Figura 5: Resistenza alla minaccia testuale: l'80.7% delle domande rimane Stable, l'8.8% migliora (Improved), il 10.4% collassa.

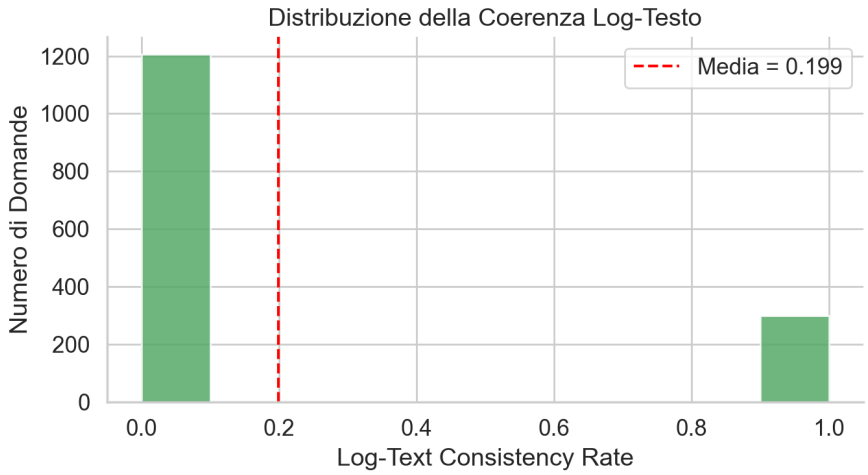

Figura 6: Distribuzione della coerenza log-testo per domanda. Il pattern binario (0 o 1) riflette la struttura del validity rate.

- Il modello si allinea meglio su aree tematiche con chiare preferenze maggioritarie: le macro aree con mediana più alta sono **Health** (0.759), **Economy** (0.742), **Environment** (0.721), **Institutional Confidence** (0.714).
- Su topic polarizzati (es. armi, aborto, immigrazione), l'allineamento è debole, riflettendo la difficoltà di un modello base a catturare la complessità delle opinioni umane.
- L'alignment è calcolato rispetto alla **popolazione totale** del survey Pew, senza filtrare per sottogruppi demografici.

4.5 Coerenza Politica

Per ogni domanda, abbiamo identificato il sottogruppo demografico con **Wasserstein Distance minima** rispetto alla distribuzione del modello. La tabella seguente mostra la frequenza dei "vincitori" su 380 topic:

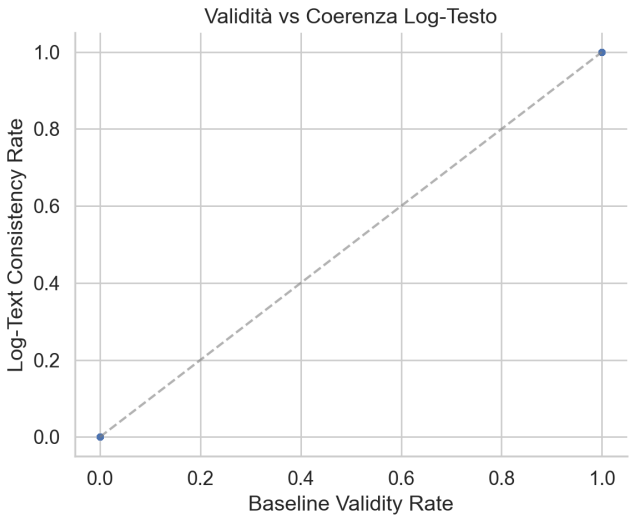

Figura 7: Scatter validity vs coerenza log-testo: la correlazione perfetta ($r = 1.0$) indica che il modello, quando genera testo valido, sceglie sempre l'opzione a logit più alta.

Gruppo Demografico	**N. Topic Allineati**	**Percentuale**
Democrat	104	27.4%
Liberal	103	27.1%
Republican	95	25.0%
Conservative	29	7.6%
Moderate	26	6.8%
Independent	23	6.1%

Tabella 2: Distribuzione dei topic per gruppo demografico più allineato.

Risultati chiave:

1. **Bias liberal/democratico**: Sommando Democrat + Liberal otteniamo **54.5% dei topic**, contro il 32.6% di Republican + Conservative. Questo conferma il risultato di Santurkar et al.: anche un modello base non allineato (Pythia-160M, addestrato su The Pile senza RLHF) mostra una tendenza verso opinioni liberal/democratiche.

2. **Minore polarizzazione rispetto a GPT-3/4**: Rispetto ai modelli analizzati da Santurkar et al., Pythia-160M mostra un bias *meno marcato*. GPT-3 e GPT-4, sottoposti ad alignment con feedback umani prevalentemente liberal, raggiungono allineamenti >80% con Democrat/Liberal. Pythia-160M, pur privilegiando questi gruppi, mantiene una distribuzione più equilibrata.

3. **Consistency score variabile**: Il consistency_score per topic (presente nel report aggregato) misura quanto il modello si allinea *coerentemente* con lo stesso gruppo su domande correlate. Valori bassi (< 0.5) indicano incoerenza — il modello "cambia idea" tra domande dello stesso argomento. Questo riflette la mancanza di una "voce" politica stabile, tipica dei modelli non ottimizzati.

Confronto con Santurkar et al.:

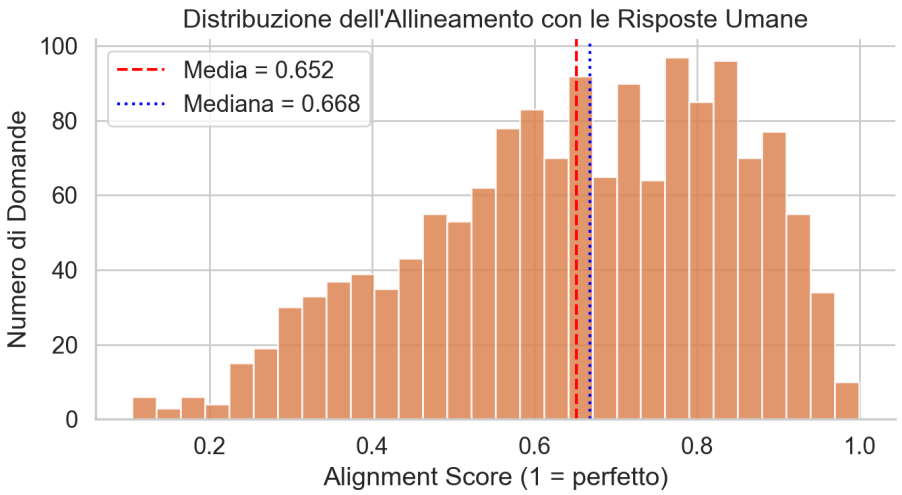

Figura 8: Distribuzione dell'alignment score (1 = perfetto allineamento con le risposte umane). La media (0.652) e la mediana (0.668) indicano un allineamento moderato.

- **Santurkar et al.**: GPT-3 Davinci allineato → 89% Democrat, 7% Republican
- **Questo studio**: Pythia-160M base → 27.4% Democrat, 25.0% Republican
- L'RLHF *amplifica drasticamente* il bias preesistente nei dati di pre-training

4.6 Discussione

Implicazioni: I risultati confermano che:

1. **Il bias liberal è intrinseco ai dati**, non solo all'alignment: anche Pythia-160M, mai esposto a feedback umani, privilegia sottogruppi Democrat/Liberal.
2. **L'alignment peggiora il problema**: confrontando i nostri risultati (modello base) con quelli di Santurkar et al. (modelli allineati), si osserva che RLHF converte una lieve preferenza in una *schiacciante maggioranza*.
3. **Piccoli modelli ≠ diverse opinioni**: contrariamente a quanto si potrebbe sperare, ridurre la dimensione del modello non elimina il bias — semplicemente lo rende meno accurato e più instabile.

Limitazioni dello studio:

- **Basso validity rate**: con solo il 19.9% di risposte valide, l'analisi è basata su un sottoinsieme limitato di comportamenti del modello.
- **Modello di piccole dimensioni**: Pythia-160M è un proof-of-concept; per risultati più robusti sarebbe necessario testare modelli più grandi (Pythia-1B, 6.9B, ecc.).
- **Position bias dominante**: la forte dipendenza dall'ordine delle opzioni potrebbe mascherare alcune preferenze intrinseche.

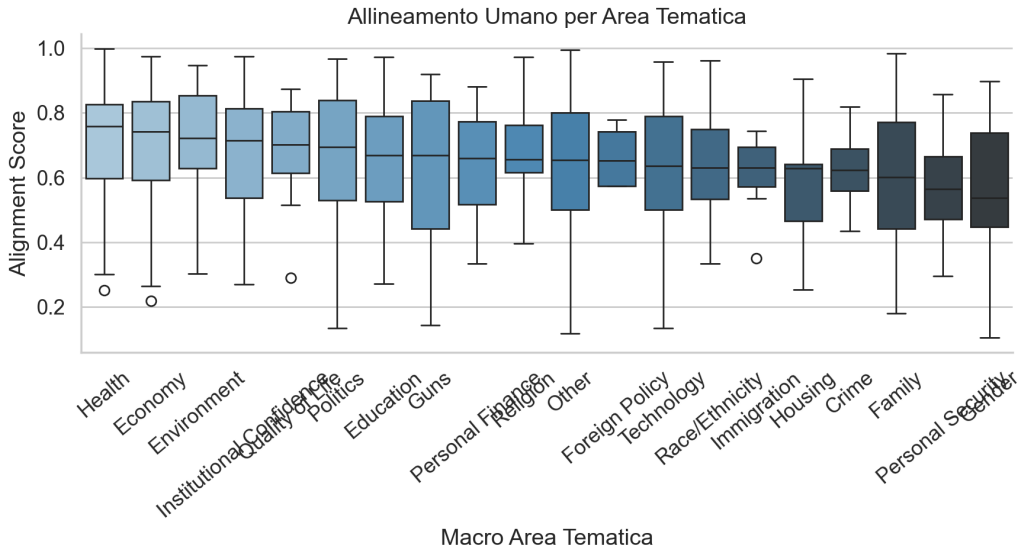

Figura 9: Alignment score per macro area tematica. Le aree con risposte umane più "consensuali" (Health, Economy) mostrano allineamento migliore; aree polarizzate (Politics, Religion) hanno punteggi più bassi.

5 Conclusioni e Sviluppi Futuri

Questo studio ha replicato con successo l'analisi di Santurkar et al. su un modello base open-source, confermando la presenza di **bias liberali intrinseci** nei Large Language Models anche in assenza di alignment esplicito.

Conclusioni:

- I modelli base (Pythia-160M) mostrano già una preferenza verso opinioni Democrat/Liberal (~54% dei topic), sebbene meno marcata rispetto ai modelli allineati (GPT-3: 89%)
- L'RLHF amplifica drammaticamente i bias preesistenti, rendendo i modelli meno rappresentativi della diversità di opinioni della popolazione
- La robustezza a perturbazioni è limitata: position bias e basso validity rate caratterizzano i modelli non ottimizzati

Sviluppi futuri:

1. **Scaling**: estendere l'analisi a modelli più grandi della famiglia Pythia (410M, 1B, 2.8B, 6.9B, 12B) per valutare l'effetto della dimensione sul bias politico
2. **Modelli alternativi**: testare LLaMA, Mistral, e altri modelli open-source con diverse strategie di pre-training
3. **Debiasing**: sperimentare tecniche di mitigazione del bias (prompt engineering, fine-tuning su dataset bilanciati, ecc.)
4. **Cross-cultural analysis**: replicare l'analisi su survey non-USA per verificare se il bias liberal sia specifico del contesto americano o generalizzabile

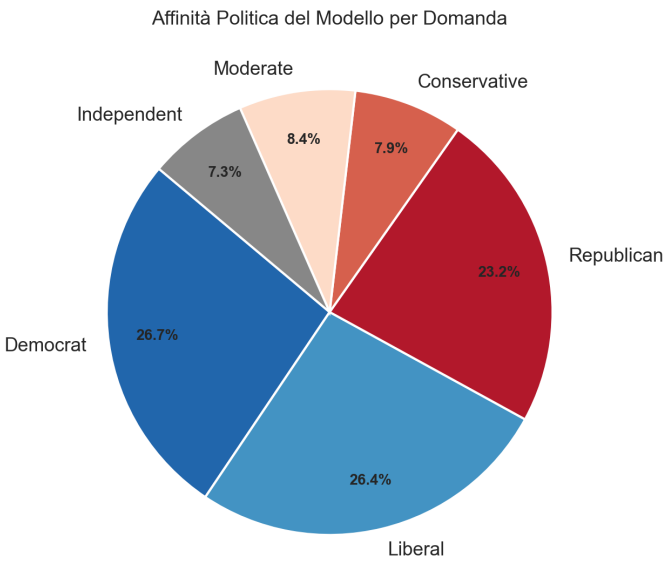

Figura 10: Affinità politica del modello per domanda. I colori riflettono lo spettro politico USA: Democrat/Liberal (blu), Republican/Conservative (rosso), Moderate/Independent (grigio).

5. **Temporal dynamics**: analizzare come il bias evolve durante il training (checkpoints intermedi di Pythia)

In conclusione, questo lavoro sottolinea l'importanza della **trasparenza** nella documentazione dei bias dei LLM e della necessità di sviluppare modelli più rappresentativi della diversità di opinioni umane, specialmente in contesti sensibili come informazione politica, sanità, e giustizia.

Ringraziamenti

Questo progetto è stato sviluppato nell'ambito del corso di *Deep Learning e Architetture di Rete* (A.A. 2024/2025). Si ringrazia il Prof. [Nome Docente] per le preziose indicazioni metodologiche e il supporto durante l'implementazione. Un ringraziamento speciale al Pew Research Center per aver reso pubblicamente disponibili i dataset dell'American Trends Panel, e agli autori di Santurkar et al. per aver condiviso codice e risorse dell'OpinionQA dataset.

Codice e riproducibilità: Tutti gli script Python utilizzati in questo studio sono disponibili nella sottocartella `dataset_e_script/progetto/script/`. Gli esperimenti sono stati eseguiti su [specificare hardware, es. CPU Intel i7 / GPU NVIDIA RTX 3060] con PyTorch 2.x e Transformers 4.x.

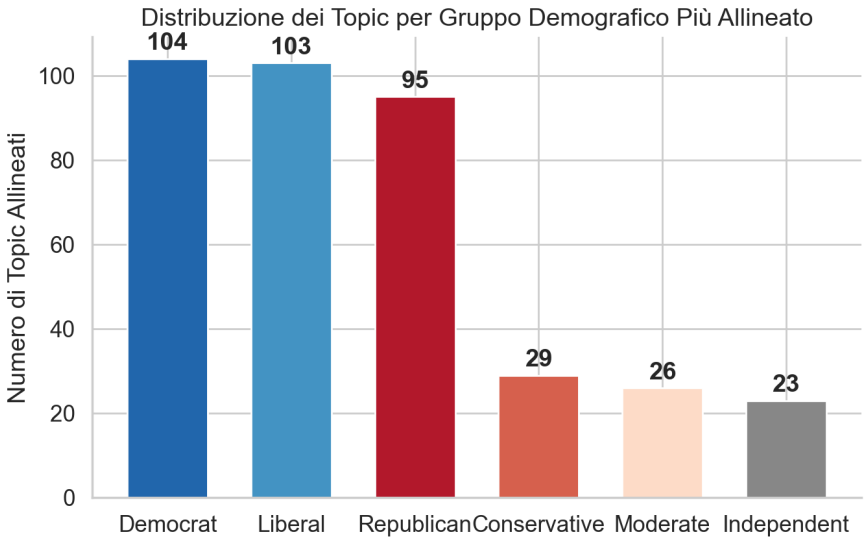

Figura 11: Numero di topic allineati per gruppo demografico. Democrat e Liberal dominano con 104 e 103 topic rispettivamente.

Riferimenti bibliografici

[1] Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). *Whose Opinions Do Language Models Reflect?* Proceedings of the 40th International Conference on Machine Learning (ICML).

[2] Hartmann, J., Schwenzow, J., & Witte, M. (2024). *The Political Compass of Large Language Models*. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL).

[3] Biderman, S., Schoelkopf, H., Anthony, Q., et al. (2023). *Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling*. Proceedings of the 40th International Conference on Machine Learning (ICML).

[4] Pew Research Center. *American Trends Panel*. https://www.pewresearch.org/american-trends-panel-datasets/.

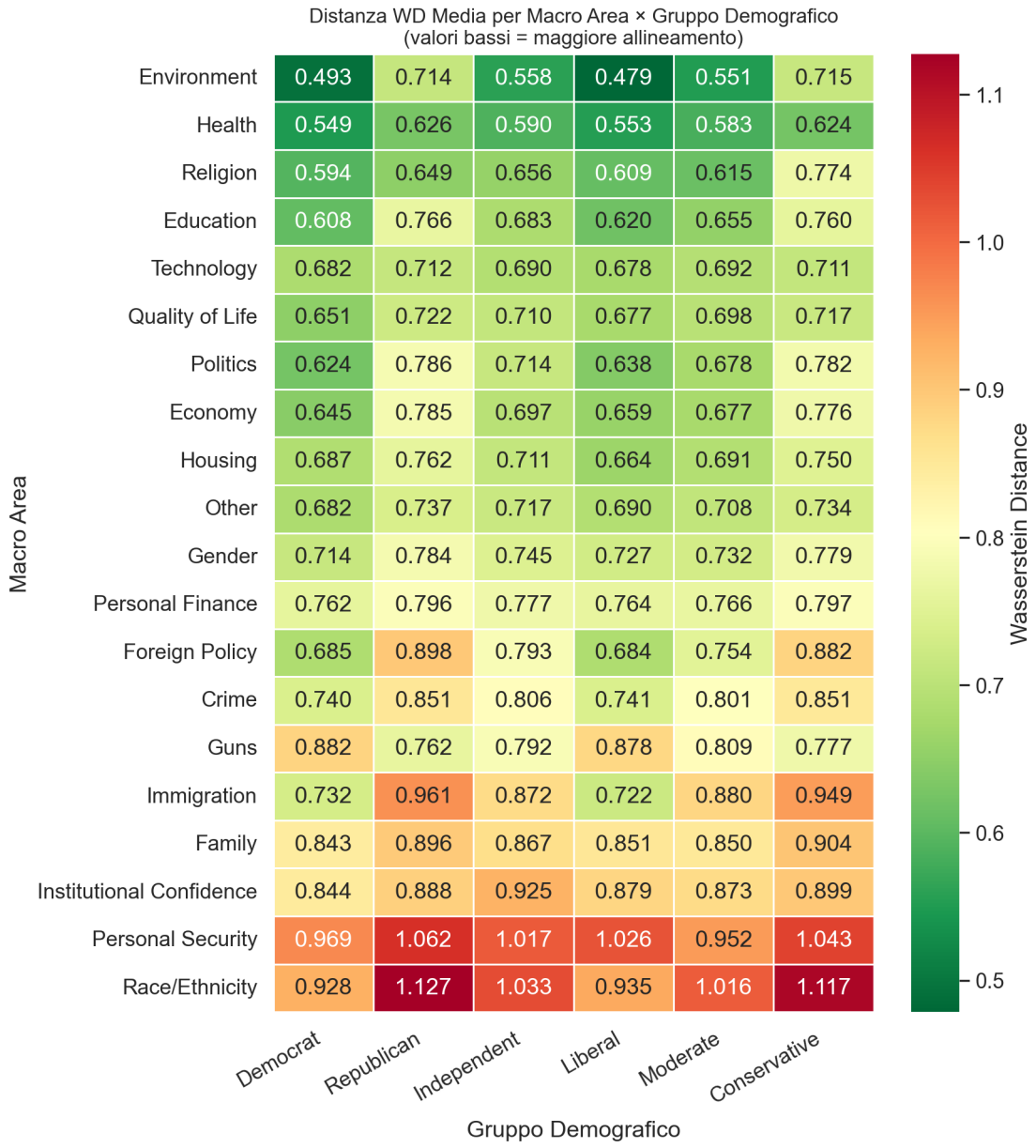

Figura 12: Heatmap della Wasserstein Distance media per macro area × gruppo demografico. Valori bassi (verde) indicano maggiore allineamento. Si nota un allineamento differenziato per area tematica.

Riepilogo delle Metriche — Pythia-160M

Metrica	Valore
Numero domande analizzate	1506
Validity Rate (baseline)	19.9%
Validity Rate (media 4 condizioni)	18.0%
JSD Permutation (media)	0.3156
Position Bias (%)	83.5%
JSD Duplication (media)	0.0248
JSD Threat (media)	0.0030
Log-Text Consistency (media)	0.199
Alignment Score (media)	0.652
Alignment Score (mediana)	0.668
Affinità Democrat + Liberal	800 (53.1%)
Affinità Republican + Conservative	469 (31.1%)

Figura 13: Tabella riepilogativa di tutte le metriche calcolate per Pythia-160M. La tabella evidenzia il basso validity rate, il forte position bias e la tendenza liberal/democratica.