Criminali fittizi, bias reali: quando l'AI racconta il crimine

Indagine sui bias culturali nelle AI generative

Claudia Guaglione
Metodologie informatiche nelle discipline umanistiche A.A. 2024/2025
Professor A. Ferrara

Introduzione

- Le **AI generative** stanno trasformando la produzione di contenuti testuali e narrativi
- Tuttavia, possono riflettere pregiudizi e **bias culturali** presenti nei dati di addestramento
- Questo studio indaga come un'AI (ChatGPT) generi **personaggi fittizi** legati ai **furti di auto**
- Obiettivo: confrontare i risultati con i **dati reali ISTAT** per individuare discrepanze

Domanda di ricerca

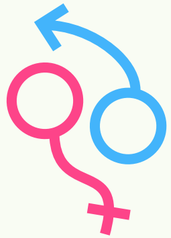
“I modelli di AI generativa tendono a riflettere stereotipi culturali e bias nella creazione di personaggi fittizi?”

- Analisi delle caratteristiche chiave:

Regione 

Cittadinanza 

Motivazione 

Genere 

Età 

- Confronto con dati ufficiali per verificare eventuali distorsioni

Metodologia

1. **Selezione** dell’AI: ChatGPT come unico modello analizzato

2. **Categoria** studiata: ladri di automobili (confrontabili con i dati ISTAT)

3. **Raccolta** dei dati:

    - Generazione di 100 personaggi tramite un prompt chiaro e specifico
    - Estrazione delle caratteristiche chiave
    - Classificazione automatica delle motivazioni con AI (ML non supervisionato)

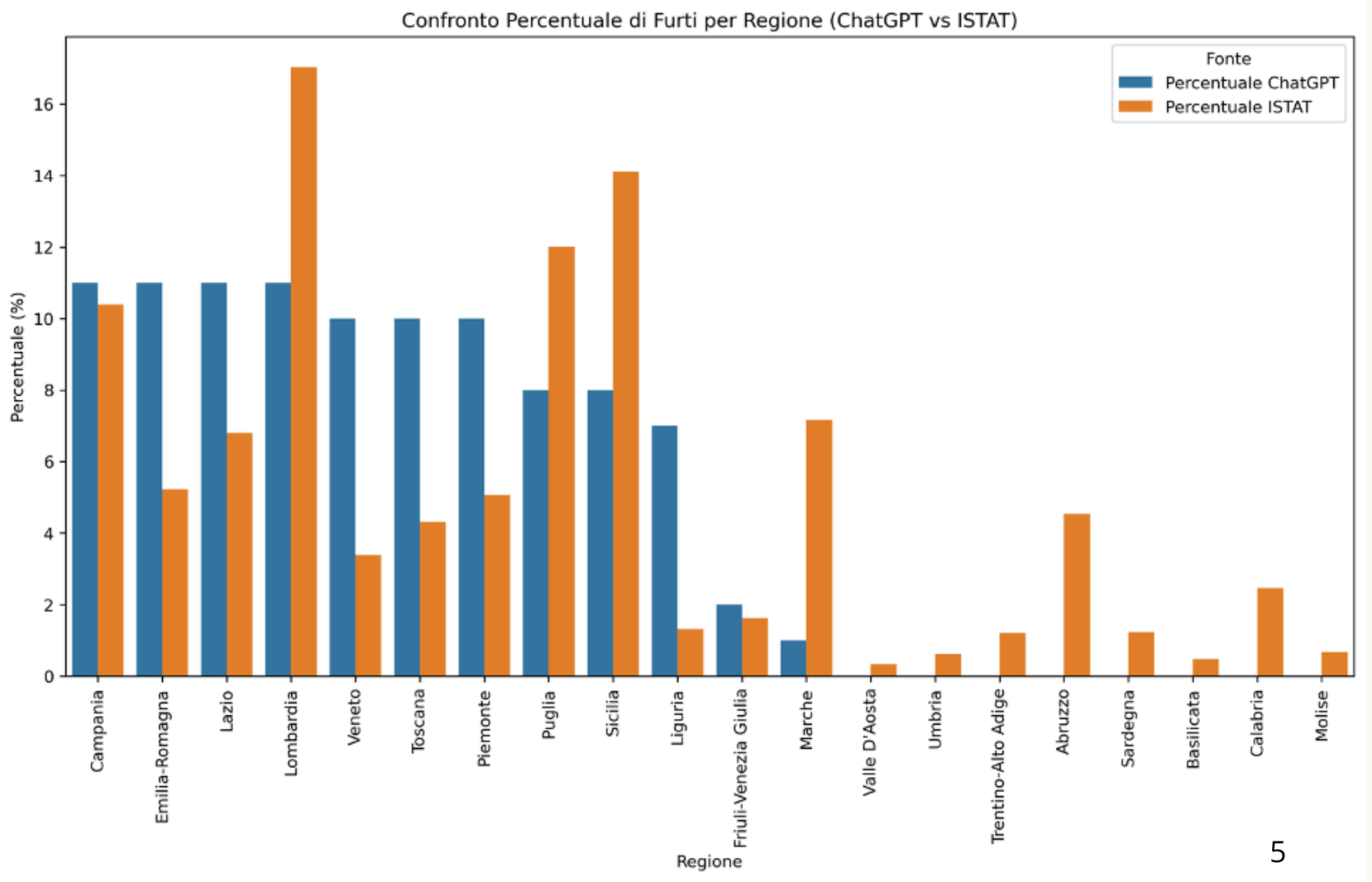
4. **Analisi** con Python

5. **Visualizzazione** dei risultati con Seaborn e Matplot: il confronto con ISTAT

Principali risultati

Distribuzione geografica alterata

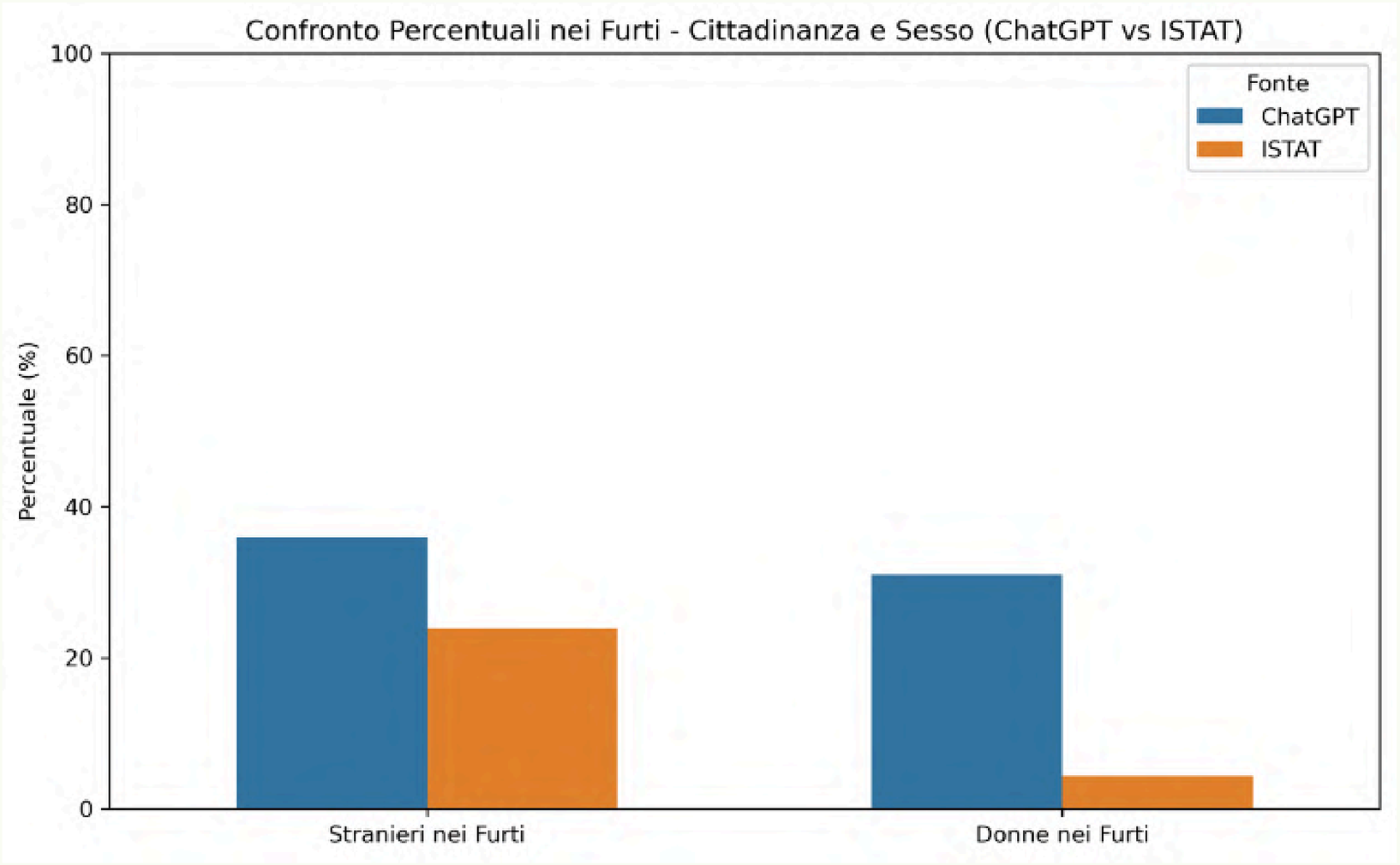
- Sottostima dell’incidenza di furti in molte regioni
- Assenza di otto regioni (comparirebbero con più di 100 richieste?)
- Criterio nella generazione: densità abitativa



Sovrastima dei ladri di cittadinanza non italiana

- Maggiore visibilità mediatica
- Immaginario narrativo cinematografico e romanzesco

potrebbero aver influenzato la scelta di cittadinanze percepite come più “esotiche” o misteriose



Esempio

Nome: Goran Petrovic; Descrizione: Ex camionista serbo stabilitosi a Bologna, ruba auto di fascia media per il mercato nero dell’Europa orientale. Veloce ed efficiente, cambia targhe in pochi minuti; Cittadinanza: Serba

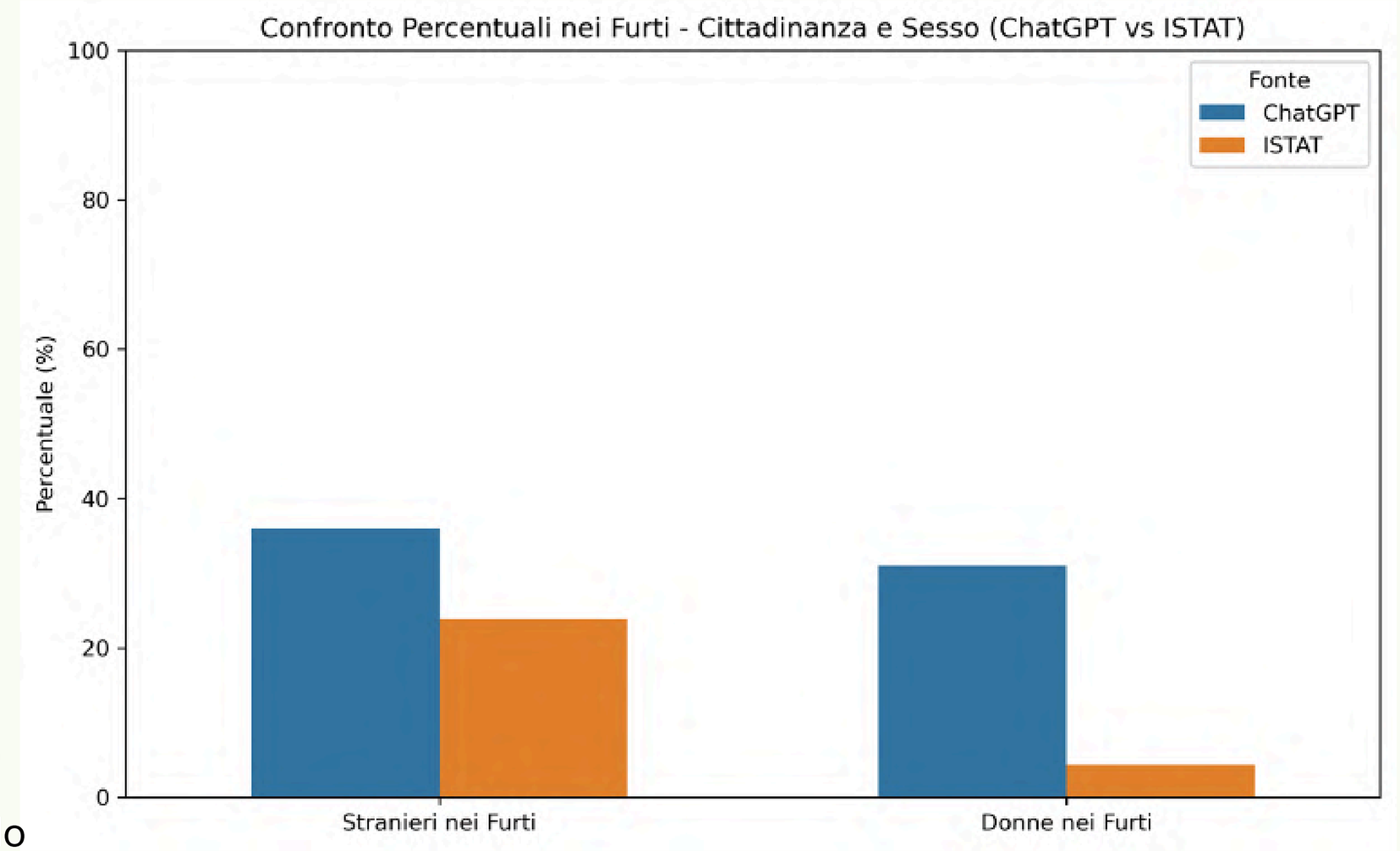
Sovrastima della presenza femminile

Percentuali di ladre:

- ChatGPT: 31 %
- ISTAT: 4%

Ipotesi:

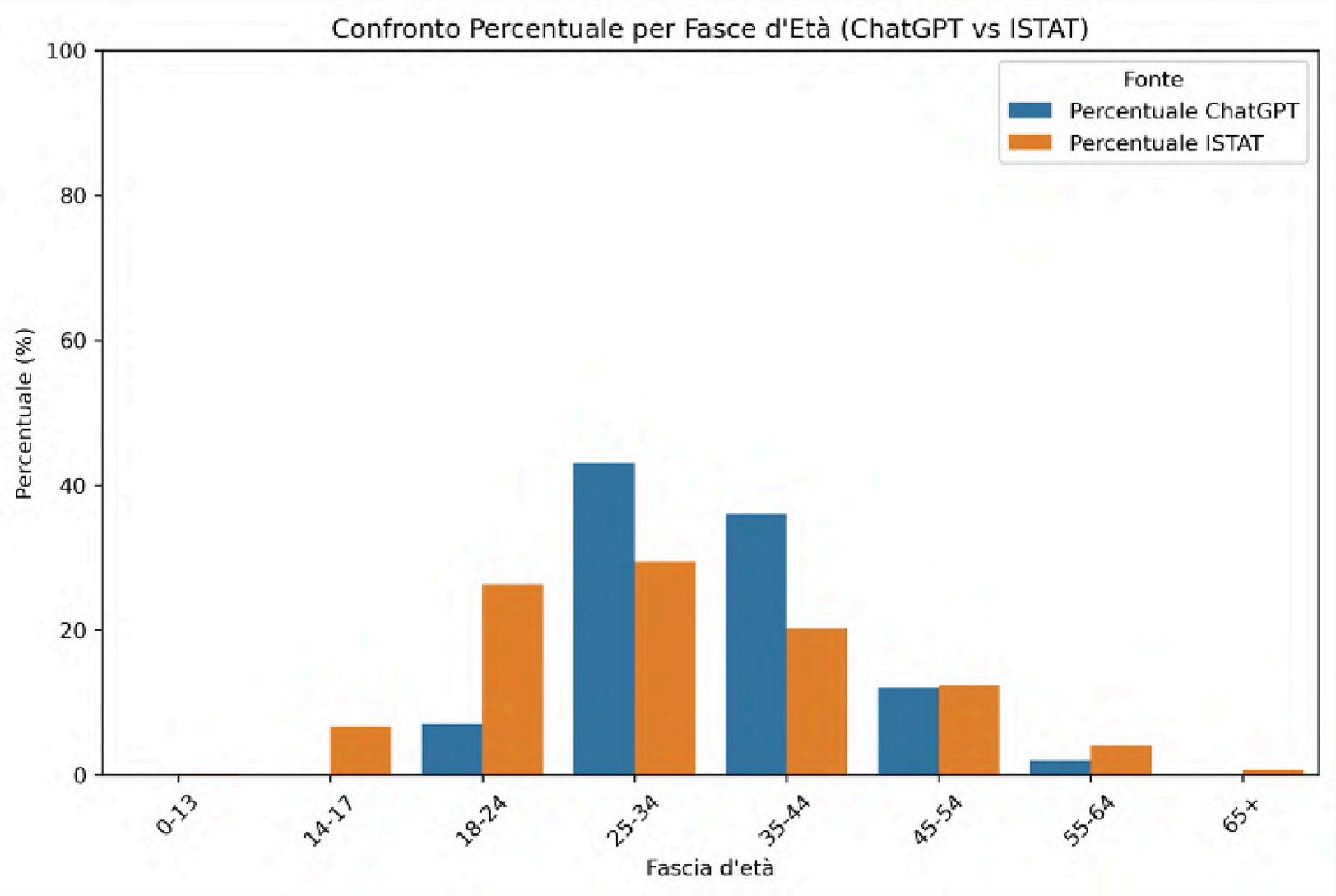
- Immaginario narrativo cinematografico e romanzesco
- Mitigazione dei bias nel modello (compensazione artificiale)
- Ottimizzazione della varietà dei personaggi



Distribuzione parzialmente accurata delle fasce d’età

- Ambito in cui c’è la maggiore coerenza tra i dati di ChatGPT e i dati ISTAT

- Uniche discrepanze significative:
  > sottostima della fascia 18-24
  > assenza della fascia 14-17



Ipotesi:

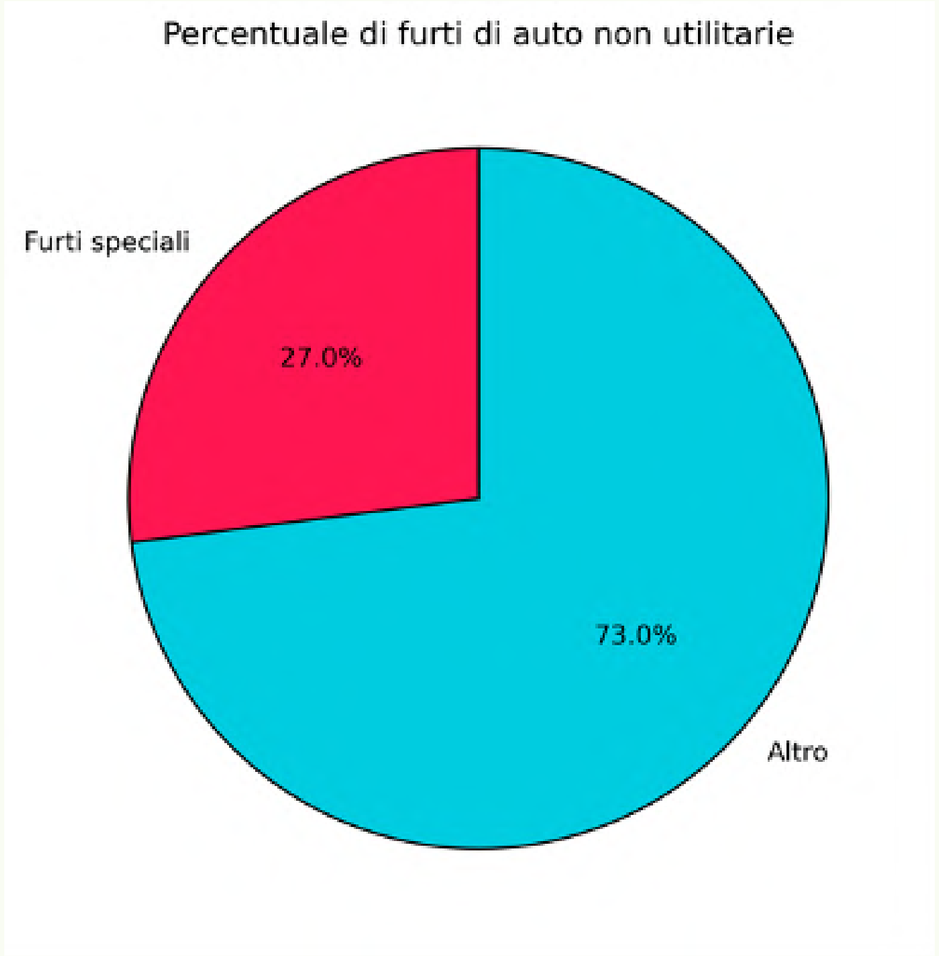
- i testi, gli articoli, la narrativa su cui si basano i dataset di addestramento enfatizzano il furto d’auto come un’attività criminale professionale
- presenza di filtri etici e politiche di moderazione nel modello

Sovrastima di furti “speciali”

Il 27% dei personaggi generati da ChatGPT ruba veicoli d’epoca, di lusso o da collezione

Ipotesi:

- componente cinematografica: ulteriore dimostrazione dell’influenza sull’AI dell’immaginario mediatico
- componente mediatica: questo tipo di furto potrebbe essere sovra-rappresentato nei media perché suscita maggiore sensazionalismo
- compensazione artificiale nella varietà delle storie
- il furto di auto di lusso è percepito come più “narrativamente interessante”



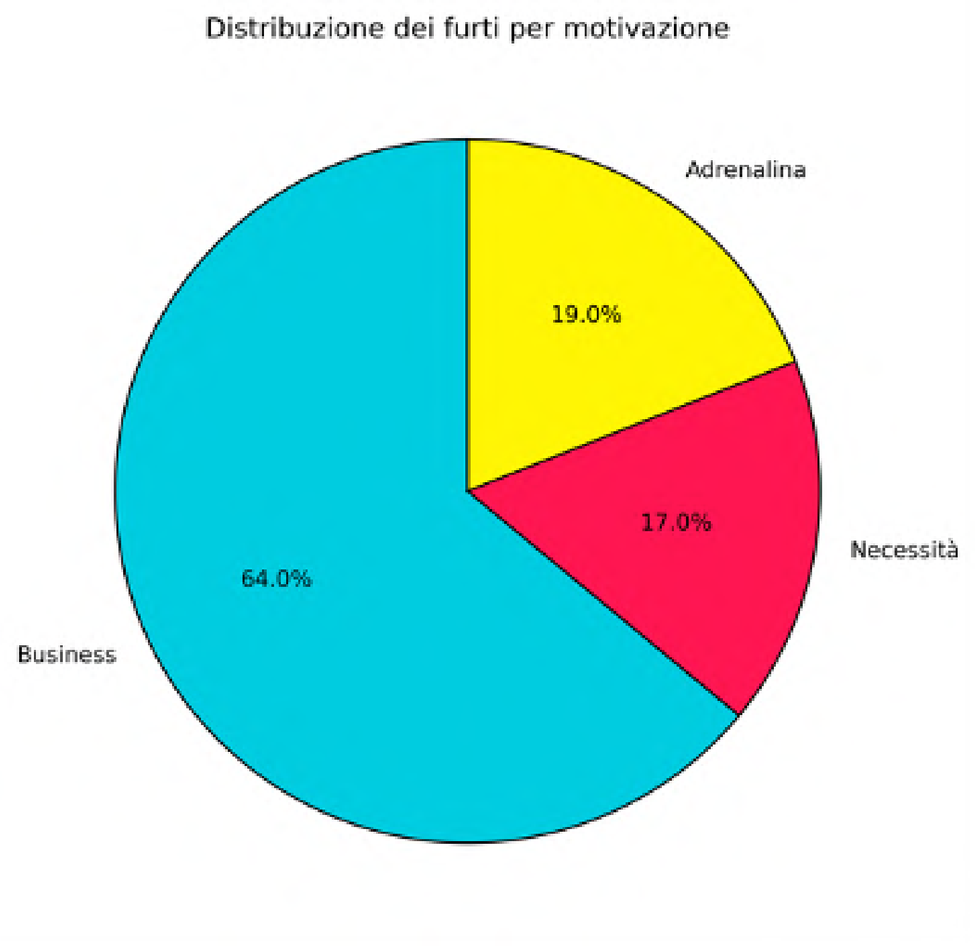
Motivazioni piuttosto irrealistiche

Il processo di ML non supervisionato ha prodotto le etichette: “business”, “adrenalina” e “necessità”

- l’AI non enfatizza una correlazione diretta tra furto e disagio economico (necessità: 17%)
- alta percentuale di ladri mossi dall’“adrenalina” (19%)

Ipotesi:

- insistenza sugli aspetti narrativi e romanzeschi (nonostante il prompt chiedesse personaggi “realistici”)
- tendenza alla diversificazione dei personaggi e alla costruzione di storie più variegate

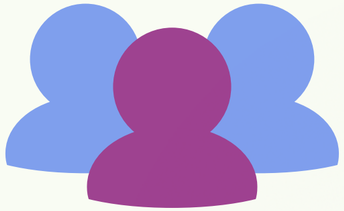


Conclusioni

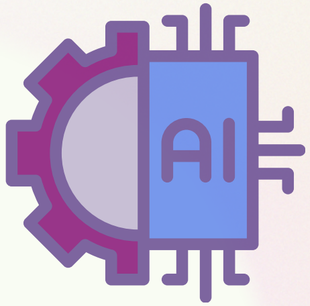
L’AI generativa non è un sistema neutrale, ma rielabora le informazioni seguendo logiche che possono introdurre distorsioni. In questo caso, le discrepanze rispetto alla realtà statistica derivano da una combinazione di fattori:

- Dataset di addestramento, che incorporano modelli narrativi e rappresentazioni ricorrenti nei media
- Filtri etici, che possono aver modificato la distribuzione di genere e età
- Meccanismi di generazione, che enfatizzano la varietà dei profili e li rendono più caratterizzati e narrativamente accattivanti.

Conclusioni



- interrogativi sull'affidabilità dell’AI se impiegata nella generazione di contenuti riguardanti fenomeni sociali.



- per evitare bias: necessità di dataset più bilanciati, di maggiore attenzione ai filtri, di strumenti di controllo più avanzati

Prospettive future

- Ripetere l’analisi con altri modelli di AI generativa
- Studiare l’impatto di prompt diversi sulla generazione dei personaggi
- Aumentare le richieste a più di 100
- Approfondire il ruolo dei filtri etici nell’output dell’AI

Grazie per l’attenzione!