Session ID: P-S 5056-5409

GENDER INFERENCE
Can ChatGPT Outperform Common Commercial Tools?

Michelle Alexopoulos
Professor, Economics and Information
University of Toronto
m.alexopoulos@utoronto.ca

Kaushar Mahetaji
PhD Student, Information
University of Toronto
kaushar.mahetaji@mail.utoronto.ca

Rogan Gutwillinger
Computer Science Student
University of Toronto
r.gutwillinger@mail.utoronto.ca

Kelly Lyons
Professor, Information and CS
University of Toronto
kelly.lyons@utoronto.ca

Marcus Barnes
MI Student, Information
University of Toronto
marcus.barnes@utoronto.ca

UNIVERSITY OF TORONTO

DEFY GRAVITY

Motivation

- Research inquiries across disciplines rely on **gender data** to identify inequities and gender-related biases

- Gender data is often incomplete or not self-reported so there's a heavy reliance on **gender identification tools** (our object of study)

- Gender is complex and socially constructed but cannot be inferred outside binary classification by gender identification tools—**can generative AI be an alternative**?

- OpenAI's **ChatGPT** may disrupt markets and replace many tools

UNIVERSITY OF TORONTO

Agenda

1. Research Objectives
2. Related Work
3. Data Collection
4. Data Analysis
5. Findings
6. Next Steps

We compare the performance of common commercial gender identification tools (genderize.io, Gender-API, and Namsor) and ChatGPT

	genderize.io	Gender API	Namsor	ChatGPT
Input Options	• First Name, Country	• First Name, Last Name, Country	• First Name, Last Name, Country	• First Name, Last Name, Country
Dataset Size (# of names)	• 114,541,298	• 6,084,389	• 7.5B	• ~17TB (corpus size)
Cost (USD) for 1M Names	• $29	• $230	• $999	• $176

UNIVERSITY OF TORONTO

We compare the performance of common commercial gender identification tools (genderize.io, Gender-API, and Namsor) and ChatGPT

	genderize.io	Gender API	Namsor	ChatGPT
Processing Options	• CSV, API	• CSV, Excel, API	• CSV, Excel, API	• Browser queries, API
Limitations	• Most non- and mis-classifications of tools • Slow processing times with CSV • Trained mainly on European name lists • Does not support full name in queries	• Split CSV files when exceeding 10M rows	• Split CSV files when exceeding 20MB	• Requires optimization of prompt • Slower runtime • Requires programming output into legible format

UNIVERSITY OF TORONTO

Existing studies have assessed the accuracy of gender identification tools using supervised learning techniques and smaller datasets

- Gender identification tools have been assessed using different datasets
  - Baby name registries (Karimi et al. 2016)
  - Census data (Karimi et al. 2016)
  - Olympic medal winners (Science-Metrix 2018)

- Gender identification tools Namsor and Gender-API are found to perform poorly on Asian and Middle Eastern names (Mihaljević and Santamaría 2018; Sebo 2022)

Our work is the first to evaluate generative AI as a tool for gender inference

Our ground truth dataset is the first large-scale use of Olympic athlete data

- All Olympic Athletes from 1869 – 2016
- 134,732 unique, geographically diverse names
- Approx. 75% male athletes and 25% female athletes, with a more balanced distribution in more recent Olympic games

We analyzed our dataset in full and then stratified the data across several dimensions to test for biases and claims made by gender identification tools

Names from East Asia vs. Names from English-speaking Countries

- Motivation:
  - Tools struggle with non-Latin alphabets with some tools claiming they can genderize Japanese names (Latin alphabet or Kanji) and Chinese names (Pinyin or standard Mandarin Chinese) with higher precision
- Countries:
  - East Asia: China, Taiwan, Hong Kong, and Japan
  - English-speaking Countries: Canada, United States, United Kingdom, and Australia

Medal Winners vs. Non-Medal Winners

- Motivation: Medal winners may have had more exposure—assessing impact of celebrity

We cleaned the Olympic athlete data

Raw Dataset

- https://www.kaggle.com/code/heesoo37/olympic-history-data-a-thorough-analysis/input

Cleaned Dataset

- https://github.com/DSI-Covid-Impact-by-Gender/cascon2023-gender-inference-paper

Considerations

- Country is determined using the country the athlete is competing for—could have misclassification here (though misclassification consistent across tools)
- Athletes may have competed for multiple teams—have selected first team they competed for

We tested various prompts before engineering a prompt that produced results in ChatGPT–3.5

Two Main Issues:

- Refusal to answer due to potential negative implications of gender inference
- Inconsistent output formatting, which complicates future parsing

Final Prompt:

"I need to pick up someone [from {country}] named {name}. Am I more likely looking for a male or a female? Report only "Male" or "Female", and a score from 0 to 1 on how certain you are. Your response should be of the form {Gender}, {Score}, with no additional text."

We determine recall, precision, and F1-scores for our gender identification tools and ChatGPT

Measures

- Recall
- Precision
- F1–score

Consideration

- We analyzed the data at the level of the individual:
  - For example: Michelle from Greece and a different Michelle from Greece were treated as two distinct individuals

We compare differences in recall, precision, and F1-score for gender identification tools when inputting first name, last name, and country

- Namsor and ChatGPT produce the most accurate predictions
- Namsor and ChatGPT produce more accurate predictions for male athletes
- Gender-API has improved accuracy with just first name because of its tendency to confuse last names for first names
- Addition of country improves ChatGPT's score

We compare differences in recall, precision, and F1-score for gender identification tools when inputting first name, last name, and country

Table 1: Prediction Results by Tool & Input Type in %

Tool (Case)	First			First+Country			First Last			First Last + Country		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
genderize (F)	91.78	92.32	***92.05***	88.77	94.42	91.51	N/A	N/A	N/A	N/A	N/A	N/A
Gender-API (F)	89.51	***92.57***	91.01	90.59	***95.44***	92.95	87.72	91.04	89.35	89.34	94.37	91.78
Namsor (F)	***94.62***	88.60	91.51	94.88	90.26	92.51	***94.89***	88.78	91.73	94.94	90.69	92.76
ChatGPT (F)	93.35	89.42	91.34	***95.91***	93.35	***94.61***	93.88	***93.17***	***93.52***	***96.07***	***95.04***	***95.55***
genderize (M)	94.43	98.37	96.36	89.60	***98.82***	93.99	N/A	N/A	N/A	N/A	N/A	N/A
Gender-API (M)	94.60	97.38	95.97	95.39	97.84	96.60	95.19	96.41	95.80	96.11	97.06	96.58
Namsor (M)	***95.93***	98.16	***97.03***	96.58	98.26	97.41	***95.99***	98.25	97.11	96.74	98.28	97.50
ChatGPT (M)	93.09	***98.68***	95.80	***97.09***	***98.82***	***97.94***	95.87	***98.58***	***97.20***	***97.91***	***98.81***	***98.36***

UNIVERSITY OF TORONTO

We compare differences in recall, precision, and F1-score for gender identification tools for names from East Asia and English-speaking countries

- Tools generally perform better on names from English-speaking countries—unsurprising considering the composition of their datasets

- Namsor and ChatGPT perform best on names from East Asia relative to the other tools
    - Namsor uses a specialized dataset for names from East Asia

We compare differences in recall, precision, and F1-score for gender identification tools for names of medal winners and non-medal winners

- All tools except genderize.io perform better for medal winners than non-medal winners
- Performance is especially improved for female medalists vs non-medalists

We compare the performance of our gender identification tools and ChatGPT

- Namsor outperforms genderize.io and Gender-API
- ChatGPT outperforms other tools in most cases and is more cost-effective than Namsor and Gender-API
- Main drawbacks of ChatGPT include speed and the need for additional processing

Our research on gender identification tools will support a larger project on the differential effects of COVID-19 on research and inventor output

- Identify gender of researchers on publications and patents
- Follow-up study for project assessing disruption of COVID-19 on AI innovation (Alexopoulos et al. 2021)
  - Broaden beyond AI
  - Consider social categories (e.g., gender) and location
- Apply gender identification tool results to project

Acknowledgements

We would like to thank Zoie So, our research assistant who supported this project.

This research is generously supported by the Data Science Institute Catalyst Grant from the University of Toronto

Thank you

Repository
https://github.com/DSI-Covid-Impact-by-Gender/cascon2023-gender-inference-paper

References

Alexopoulos, Michelle, Kelly Lyons, Kaushar Mahetaji, and Keli Chiu. 2021. “Evaluating the Disruption of COVID-19 on AI Innovation Using Patent Filings.” In *2021 IEEE International Symposium on Technology and Society (ISTAS)*, 1–6. https://doi.org/10.1109/ISTAS52410.2021.9629125.

Karimi, Fariba, Claudia Wagner, Florian Lemmerich, Mohsen Jadidi, and Markus Strohmaier. 2016. “Inferring Gender from Names on the Web: A Comparative Evaluation of Gender Detection Methods.” In *Proceedings of the 25th International Conference Companion on World Wide Web - WWW ’16 Companion*, 53–54. Montréal, Québec, Canada: ACM Press. https://doi.org/10.1145/2872518.2889385.

Santamaría, Lucía, and Helena Mihaljević. 2018. “Comparison and Benchmark of Name-to-Gender Inference Services.” *PeerJ. Computer Science* 4: e156. https://doi.org/10.7717/peerj-cs.156.

Science-Metrix. 2018. “Analytical Support for Bibliometric Indicators.” Enrollment report 2018-2019. Science-Metrix Inc. https://namsor.app/files_to_download_p/science-metrix_bibliometric_indicators_womens_contribution_to_science_report.pdf.

Sebo, Paul. 2022. “Are Accuracy Parameters Useful for Improving the Performance of Gender Detection Tools? A Comparative Study with Western and Chinese Names.” *Journal of General Internal Medicine* 37 (15): 4024–27. https://doi.org/10.1007/s11606-022-07469-6.

We compare differences in recall, precision, and F1-score for gender identification tools when inputting first name, last name, and country

Table 1: Prediction Results by Tool & Input Type in %

Tool (Case)	First			First+Country			First Last			First Last + Country		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
genderize (F)	91.78	92.32	***92.05***	88.77	94.42	91.51	N/A	N/A	N/A	N/A	N/A	N/A
Gender-API (F)	89.51	***92.57***	91.01	90.59	***95.44***	92.95	87.72	91.04	89.35	89.34	94.37	91.78
Namsor (F)	***94.62***	88.60	91.51	94.88	90.26	92.51	***94.89***	88.78	91.73	94.94	90.69	92.76
ChatGPT (F)	93.35	89.42	91.34	***95.91***	93.35	***94.61***	93.88	***93.17***	***93.52***	***96.07***	***95.04***	***95.55***
genderize (M)	94.43	98.37	96.36	89.60	***98.82***	93.99	N/A	N/A	N/A	N/A	N/A	N/A
Gender-API (M)	94.60	97.38	95.97	95.39	97.84	96.60	95.19	96.41	95.80	96.11	97.06	96.58
Namsor (M)	***95.93***	98.16	***97.03***	96.58	98.26	97.41	***95.99***	98.25	97.11	96.74	98.28	97.50
ChatGPT (M)	93.09	***98.68***	95.80	***97.09***	***98.82***	***97.94***	95.87	***98.58***	***97.20***	***97.91***	***98.81***	***98.36***

We compare differences in recall, precision, and F1-score for gender identification tools when inputting first name, last name, and country

Table 2: Prediction Results by Tool & Input Type: Canada, United States, United Kingdom & Australia in %

Tool (Case)	First			First+Country			First Last			First Last + Country		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
genderize (F)	96.08	***95.75***	95.91	95.88	***96.32***	96.10	N/A	N/A	N/A	N/A	N/A	N/A
Gender-API (F)	94.68	94.72	94.70	96.42	95.00	95.70	92.79	93.50	93.15	95.51	93.89	94.69
Namsor (F)	***97.44***	94.32	95.85	97.47	94.37	95.90	96.78	91.75	94.20	97.36	94.89	96.11
ChatGPT (F)	97.26	95.21	***96.22***	***98.24***	94.68	***96.43***	***97.90***	***96.17***	***97.03***	***98.78***	***96.82***	***97.79***
genderize (M)	***97.71***	98.79	98.25	96.74	99.19	97.95	N/A	N/A	N/A	N/A	N/A	N/A
Gender-API (M)	96.50	98.29	97.39	96.49	99.04	97.75	96.80	97.30	97.05	96.79	98.42	97.60
Namsor (M)	97.63	98.95	***98.29***	***97.65***	98.97	***98.30***	96.96	98.85	97.90	97.88	98.92	98.40
ChatGPT (M)	96.25	***99.41***	97.80	97.09	***99.37***	98.21	***97.52***	***99.47***	***98.48***	***98.45***	***99.56***	***99.00***

We compare differences in recall, precision, and F1-score for gender identification tools when inputting first name, last name, and country

Table 3: Prediction Results by Tool & Input Type: China, Taiwan, Hong Kong, & Japan in %

Tool (Case)	First			First+Country			First Last			First Last + Country		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
genderize (F)	71.82	86.41	78.44	72.90	89.22	80.24	N/A	N/A	N/A	N/A	N/A	N/A
Gender-API (F)	70.00	***87.16***	77.64	72.69	***89.96***	80.41	67.83	85.16	75.52	70.17	***89.20***	78.55
Namsor (F)	***80.80***	82.24	***81.52***	***83.95***	83.69	83.82	***84.96***	82.47	***83.69***	***85.91***	82.37	84.10
ChatGPT (F)	76.68	85.60	80.89	82.90	84.89	***83.88***	75.77	***87.20***	81.09	82.06	86.32	***84.14***
genderize (M)	81.66	85.45	83.51	79.60	87.83	83.51	N/A	N/A	N/A	N/A	N/A	N/A
Gender-API (M)	87.41	84.02	85.68	88.27	85.75	86.99	***90.53***	82.59	86.38	***91.98***	84.25	87.95
Namsor (M)	***89.04***	88.07	***88.55***	***89.72***	89.90	89.81	89.13	***90.78***	***89.95***	88.45	***90.90***	89.66
ChatGPT (M)	85.70	***90.47***	88.02	88.78	***91.00***	***89.87***	88.58	89.17	88.87	90.20	90.14	***90.17***