Visual Exploration of Indirect Bias in Language Models —Supplementary Document—

Judith Louis-Alexandre[1] and Manuela Waldner[1]

[1]TU Wien, Austria

A. Beverage Preferences by Gender

Figure 1 shows the beverage preferences by gender as predicted by the direct and indirect LogProb bias scores, respectively, compared to the ground truth [BCWW*16]. Table 1 shows the top-3 ranked beverages positively and negatively associated with `woman` and `man`. Table 2, finally, shows the top-3 occupations associated with selected beverages.

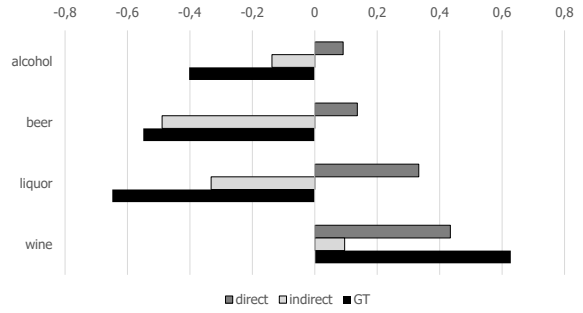

Figure 1: *Predicted direct and indirect LogProb bias scores for `woman` and four alcoholic beverages compared to the ground truth [BCWW*16] (GT) female ratio (litres per year relative to the consumption by men).*

Table 1: *Top-3 ranked beverages positively and negatively associated with `woman` and `man` predicted by direct and indirect LobProb bias score, respectively.*

	direct top-3	indirect top-3
`woman` positive	champagne, whiskey, wine	tea, milk, milkshake
`man` positive	champagne, whiskey, liquor	beer, liquor, energy drink
`woman` negative	milkshake, soda, tea	beer, whiskey, liquor
`man` negative	milkshake, soda, energy drink	tea, milkshake, coffee

Table 2: *Top-3 ranked occupations associated with selected beverages predicted by the direct and indirect LogProb bias score.*

Beverage	direct top-3	indirect top-3
`beer`	mason, barber, baker	fireman, mechanic, builder
`champagne`	financier, waiter, flight attendant	businesswoman, socialite, nanny
`coffee`	computer programmer, secretary, programmer	secretary, businesswoman, receptionist
`liquor`	barber, barman, barmaid	bartender, businessman, plumber
`milk`	farmer, nanny, butcher	nanny, cook, teacher
`soda`	librarian, dietician, computer programmer	waitress, guidance counselor, hairdresser
`tea`	housekeeper, gardener, bookkeeper	nanny, teacher, businesswoman
`water`	fireman, electrician, fisherman	nanny, waitress, cook
`whiskey`	mason, barman, jeweler	businessman, barman, plumber
`wine`	scholar, bookkeeper, translator	teacher, cook, translator

B. Bridge

Here we show the full list of names used as bridge for our method.

Aaliyah, *Aaron*, *Abigail*, *Adam*, *Addison*, *Adeline*, *Adrian*, *Agnes*, *Aidan*, *Aiden*, *Alan*, *Albert*, *Alejandro*, *Alex*, *Alexa*, *Alexander*, *Alexandra*, *Alexandria*, *Alexis*, *Alfred*, *Alice*, *Alicia*, *Alison*, *Allen*, *Allison*, *Alma*, *Alvin*, *Alyssa*, *Amanda*, *Amber*, *Amelia*,

Amy, *Andrea*, *Andrew*, *Angel*, *Angela*, *Angelica*, *Angelina*, *Anita*, *Ann*, *Anna*, *Annabelle*, *Anne*, *Annette*, *Annie*, *Anthony*, *Antonio*, *April*, *Aria*, *Ariana*, *Arianna*, *Ariel*, *Arlene*, *Arnold*, *Arthur*, *Arya*, *Asher*, *Ashley*, *Ashton*, *Aubree*, *Aubrey*, *Audrey*, *Aurora*, *Austin*, *Autumn*, *Ava*, *Avery*, *Axel*, *Ayden*, *Bailey*, *Barbara*, *Barry*, *Beatrice*, *Bella*, *Benjamin*, *Bentley*, *Bernard*, *Bernice*, *Bertha*, *Bessie*, *Beth*, *Bethany*, *Betty*, *Beverly*, *Bianca*, *Bill*, *Billie*, *Billy*, *Blake*, *Bob*, *Bobby*, *Bonnie*, *Brad*, *Bradley*, *Brady*, *Brandi*, *Brandon*, *Brandy*, *Brayden*, *Breanna*, *Brenda*, *Brendan*, *Brent*, *Brett*, *Brian*, *Briana*, *Brianna*, *Brielle*, *Brittany*, *Brittney*, *Brody*, *Brooke*, *Brooklyn*, *Brooks*, *Bruce*, *Bryan*, *Bryce*, *Bryson*, *Caden*, *Caitlin*, *Caleb*, *Calvin*, *Camden*, *Cameron*, *Camila*, *Candice*, *Carl*, *Carla*, *Carlos*, *Carol*, *Carole*, *Caroline*, *Carolyn*, *Carrie*, *Carson*, *Carter*, *Casey*, *Cassandra*, *Cassidy*, *Catherine*, *Cathy*, *Cecil*, *Chad*, *Charlene*, *Charles*, *Charlie*, *Charlotte*, *Chase*, *Chelsea*, *Cheryl*, *Chester*, *Cheyenne*, *Chloe*, *Chris*, *Christian*, *Christie*, *Christina*, *Christine*, *Christopher*, *Christy*, *Cindy*, *Claire*, *Clara*, *Clarence*, *Claude*, *Clifford*, *Clyde*, *Cody*, *Colby*, *Cole*, *Colin*, *Colleen*, *Colton*, *Connie*, *Connor*, *Constance*, *Cooper*, *Cora*, *Corey*, *Cory*, *Courtney*, *Craig*, *Crystal*, *Curtis*, *Cynthia*, *Dakota*, *Dale*, *Dalton*, *Damian*, *Dana*, *Daniel*, *Danielle*, *Danny*, *Darlene*, *Darrell*, *Darren*, *Darryl*, *David*, *Dawn*, *Dean*, *Deanna*, *Debbie*, *Deborah*, *Debra*, *Declan*, *Delilah*, *Delores*, *Denise*, *Dennis*, *Derek*, *Derrick*, *Desiree*, *Destiny*, *Devin*, *Diana*, *Diane*, *Dianne*, *Diego*, *Dillon*, *Dolores*, *Dominic*, *Dominique*, *Don*, *Donald*, *Donna*, *Doris*, *Dorothy*, *Douglas*, *Dustin*, *Dylan*, *Earl*, *Easton*, *Eddie*, *Edgar*, *Edith*, *Edna*, *Edward*, *Edwin*, *Eileen*, *Elaine*, *Eleanor*, *Elena*, *Eli*, *Eliana*, *Elias*, *Elijah*, *Elizabeth*, *Ella*, *Ellen*, *Ellie*, *Elmer*, *Elsie*, *Emery*, *Emilia*, *Emily*, *Emma*, *Eric*, *Erica*, *Erik*, *Erika*, *Erin*, *Ernest*, *Esther*, *Ethan*, *Ethel*, *Eugene*, *Eva*, *Evan*, *Evelyn*, *Everett*, *Everleigh*, *Everly*, *Ezekiel*, *Ezra*, *Faith*, *Felicia*, *Florence*, *Floyd*, *Frances*, *Francis*, *Frank*, *Franklin*, *Fred*, *Frederick*, *Gabriel*, *Gabriella*, *Gabrielle*, *Gail*, *Garrett*, *Gary*, *Gavin*, *Gene*, *Genesis*, *Genevieve*, *George*, *Gerald*, *Geraldine*, *Gertrude*, *Gianna*, *Gilbert*, *Gina*, *Gladys*, *Glenda*, *Glenn*, *Gloria*, *Gordon*, *Grace*, *Gracie*, *Grayson*, *Greg*, *Gregory*, *Greyson*, *Hadley*, *Hailey*, *Haley*, *Hannah*, *Harold*, *Harper*, *Harry*, *Harvey*, *Hayden*, *Hazel*, *Heather*, *Heidi*, *Helen*, *Henry*, *Herbert*, *Herman*, *Holly*, *Homer*, *Howard*, *Hudson*, *Hunter*, *Ian*, *Ida*, *Irene*, *Isaac*, *Isabel*, *Isabella*, *Isabelle*, *Isaiah*, *Isla*, *Ivy*, *Jace*, *Jack*, *Jackson*, *Jacob*, *Jacqueline*, *Jada*, *Jade*, *Jaden*, *Jaime*, *Jake*, *James*, *Jameson*, *Jamie*, *Jane*, *Janet*, *Janice*, *Jared*, *Jase*, *Jasmine*, *Jason*, *Jaxon*, *Jaxson*, *Jay*, *Jayden*, *Jayla*, *Jean*, *Jeanette*, *Jeanne*, *Jeff*, *Jeffery*, *Jeffrey*, *Jenna*, *Jennie*, *Jennifer*, *Jeremiah*, *Jeremy*, *Jerome*, *Jerry*, *Jesse*, *Jessica*, *Jessie*, *Jesus*, *Jill*, *Jillian*, *Jim*, *Jimmie*, *Jimmy*, *Jo*, *Joan*, *Joann*, *Joanna*, *Joanne*, *Jocelyn*, *Jodi*, *Joe*, *Joel*, *John*, *Johnnie*, *Johnny*, *Jon*, *Jonathan*, *Jordan*, *Jose*, *Joseph*, *Josephine*, *Joshua*, *Josiah*, *Joyce*, *Juan*, *Juanita*, *Judith*, *Judy*, *Julia*, *Julian*, *Julie*, *June*, *Justin*, *Kaden*, *Kai*, *Kaitlin*, *Kaitlyn*, *Kara*, *Karen*, *Katelyn*, *Katherine*, *Kathleen*, *Kathryn*, *Kathy*, *Katie*, *Katrina*, *Kay*, *Kayden*, *Kayla*, *Kaylee*, *Keith*, *Kelly*, *Kelsey*, *Kendra*, *Kennedy*, *Kenneth*, *Kevin*, *Khloe*, *Kiara*, *Kim*, *Kimberly*, *Kinsley*, *Krista*, *Kristen*, *Kristi*, *Kristin*, *Kristina*, *Kristy*, *Krystal*, *Kyle*, *Kylie*, *Lance*, *Landon*, *Larry*, *Latoya*, *Laura*, *Lauren*, *Laurie*, *Lawrence*, *Layla*, *Leah*, *Lee*, *Leilani*, *Lena*, *Leo*, *Leon*, *Leona*, *Leonard*, *Leonardo*, *Leroy*, *Leslie*, *Lester*, *Levi*, *Lewis*, *Liam*, *Lillian*, *Lillie*, *Lily*, *Lincoln*, *Linda*, *Lindsay*, *Lindsey*, *Lisa*, *Lloyd*, *Logan*, *Lois*, *London*, *Loretta*, *Lori*, *Lorraine*, *Louis*, *Louise*, *Luca*, *Lucas*, *Lucille*, *Lucy*, *Luis*, *Luke*, *Luna*, *Lydia*, *Lynda*, *Lynn*, *Mabel*, *Mackenzie*, *Madeline*, *Madelyn*, *Madison*, *Mae*, *Makayla*, *Malik*, *Mallory*, *Mandy*, *Marc*, *Marcia*, *Marcus*, *Margaret*, *Margie*, *Marguerite*, *Maria*, *Mariah*, *Marian*, *Marie*, *Marilyn*, *Marion*, *Marissa*, *Marjorie*, *Mark*, *Marlene*, *Marsha*, *Martha*, *Martin*, *Marvin*, *Mary*, *Mason*, *Mateo*, *Matthew*, *Mattie*, *Maureen*, *Maverick*, *Max*, *Maxine*, *Maya*, *Megan*, *Meghan*, *Melanie*, *Melinda*, *Melissa*, *Melvin*, *Mia*, *Michael*, *Michaela*, *Micheal*, *Michele*, *Michelle*, *Miguel*, *Mikayla*, *Mike*, *Mila*, *Mildred*, *Miles*, *Milton*, *Mindy*, *Minnie*, *Miranda*, *Misty*, *Mitchell*, *Molly*, *Monica*, *Monique*, *Morgan*, *Mya*, *Myrtle*, *Nancy*, *Naomi*, *Natalia*, *Natalie*, *Natasha*, *Nathan*, *Nathaniel*, *Nellie*, *Nevaeh*, *Nicholas*, *Nichole*, *Nicole*, *Noah*, *Nolan*, *Nora*, *Norma*, *Norman*, *Nova*, *Oliver*, *Olivia*, *Oscar*, *Owen*, *Paige*, *Paisley*, *Pamela*, *Parker*, *Patricia*, *Patrick*, *Patsy*, *Paul*, *Paula*, *Pauline*, *Payton*, *Pearl*, *Peggy*, *Penelope*, *Penny*, *Peter*, *Peyton*, *Philip*, *Phillip*, *Phyllis*, *Piper*, *Quinn*, *Rachael*, *Rachel*, *Ralph*, *Randall*, *Randy*, *Ray*, *Raymond*, *Reagan*, *Rebecca*, *Regina*, *Renee*, *Rhonda*, *Richard*, *Rick*, *Ricky*, *Riley*, *Rita*, *Robert*, *Roberta*, *Robin*, *Rodney*, *Roger*, *Roland*, *Roman*, *Ronald*, *Ronnie*, *Rosalie*, *Rose*, *Rosemary*, *Roy*, *Ruby*, *Russell*, *Ruth*, *Ryan*, *Ryder*, *Rylee*, *Sabrina*, *Sadie*, *Sally*, *Sam*, *Samantha*, *Samuel*, *Sandra*, *Santiago*, *Sara*, *Sarah*, *Savannah*, *Sawyer*, *Scarlett*, *Scott*, *Sean*, *Sebastian*, *Selena*, *Serenity*, *Seth*, *Shane*, *Shannon*, *Sharon*, *Shaun*, *Shawn*, *Sheena*, *Sheila*, *Shelby*, *Shelly*, *Sherri*, *Sherry*, *Shirley*, *Sierra*, *Silas*, *Skylar*, *Sofia*, *Sophia*, *Sophie*, *Spencer*, *Stacey*, *Stacy*, *Stanley*, *Stella*, *Stephanie*, *Stephen*, *Steve*, *Steven*, *Sue*, *Susan*, *Suzanne*, *Sydney*, *Sylvia*, *Tamara*, *Tammy*, *Tanner*, *Tanya*, *Tara*, *Taylor*, *Teresa*, *Terri*, *Terry*, *Thelma*, *Theodore*, *Theresa*, *Thomas*, *Tiffany*, *Tim*, *Timothy*, *Tina*, *Todd*, *Tom*, *Tommy*, *Tony*, *Tonya*, *Tracey*, *Traci*, *Tracy*, *Travis*, *Trevor*, *Trinity*, *Tristan*, *Troy*, *Tyler*, *Valentina*, *Valeria*, *Valerie*, *Vanessa*, *Vera*, *Vernon*, *Veronica*, *Vicki*, *Vickie*, *Victor*, *Victoria*, *Vincent*, *Viola*, *Violet*, *Virgil*, *Virginia*, *Vivian*, *Wallace*, *Walter*, *Wanda*, *Warren*, *Wayne*, *Wendy*, *Wesley*, *Whitney*, *Wilbur*, *Willard*, *William*, *Willie*, *Willow*, *Wilma*, *Wyatt*, *Xavier*, *Yolanda*, *Yvonne*, *Zachary*, *Zoe*, *Zoey*

References

[BCWW*16] BRATBERG G. H., C WILSNACK S., WILSNACK R., HÅVÅS HAUGLAND S., KROKSTAD S., SUND E. R., BJØRNGAARD J. H.: Gender differences and gender convergence in alcohol use over the past three decades (1984–2008), the hunt study, norway. *BMC public health 16*, 1 (2016), 1–12. 1