Abstract

The goal of question answering (QA) is to answer any question. However, major QA datasets have skewed distributions over gender, profession, and nationality. Despite that skew, model accuracy analysis reveals little evidence that accuracy is lower for people based on gender or nationality; instead, there is more variation on professions (question topic). But QA’s lack of representation could itself hide evidence of bias, necessitating QA datasets that better represent global diversity.

1 Introduction

Question answering (QA) systems have impressive recent victories—beating trivia masters (Ferrucci et al., 2010) and superhuman reading (Najberg, 2018)—but these triumphs hold only if they generalize; QA systems should be able to answer questions even if they do not look like training examples. While other work (Section 4) focuses on demographic representation in NLP resources, our focus is how well QA models generalize across demographic subsets.

After mapping mentions to a knowledge base (Section 2), we show existing QA datasets lack diversity in the gender and national origin of the people mentioned: English-language QA datasets mostly ask about US men from a few professions (Section 2.2). This is problematic because most English speakers (and users of English QA systems) are not from the US or UK. Moreover, multilingual QA datasets are often translated from English datasets (Lewis et al., 2020; Artetxe et al., 2019). However, no work has verified that QA systems generalize to infrequent demographic groups.

Section 3 investigates whether statistical tests reveal patterns on demographic subgroups. Despite skewed distributions, accuracy is not correlated with gender or nationality, though it is with professional field. For instance, Natural Questions (Kwiatkowski et al., 2019, NQ) systems do well with entertainers but poorly with scientists, which are handled well in TriviaQA. However, absence of evidence is not evidence of absence (Section 5), and existing QA datasets are not yet diverse enough to vet QA’s generalization.

2 Mapping Questions to Entities

We analyze four QA tasks: NQ,1 SQuAD (Rajpurkar et al., 2016), QB (Boyd-Graber et al., 2012) and TriviaQA (Joshi et al., 2017). Google CLOUD-NL2 finds and links entity mentions in QA examples.3

2.1 Focus on People

Many entities appear in examples (Table 1) but people form a majority in our QA tasks (except SQuAD). Existing work in AI fairness focuses on disparate impacts on people, and model behaviors are prone to harm especially when it comes to people; hence, our primary intent is to understand how demographic characteristics of “people” correlate with model correctness.

The people asked about in a question can be in

| Entity Type | NQ Train | Dev | QB Train | Dev | SQuAD Train | Dev | TriviaQA Train | Dev |
|-------------|----------|-----|----------|-----|-------------|-----|---------------|-----|
| Person      | 32.14    | 32.27| 58.49    | 57.18| 14.56       | 9.56| 40.89         | 40.69|
| No entity   | 21.97    | 22.61| 15.88    | 20.13| 37.27       | 47.95| 14.85         | 14.93|
| Location    | 28.42    | 27.82| 34.50    | 35.11| 33.52       | 29.34| 44.52         | 44.08|
| Work of art | 17.31    | 16.19| 27.98    | 28.16| 3.15        | 1.15| 17.53         | 17.76|
| Other       | 2.66     | 2.70 | 9.83     | 9.88 | 5.08        | 3.10| 14.38         | 14.64|
| Organization| 17.73    | 18.05| 20.37    | 17.78| 21.34       | 20.14| 22.15         | 21.14|
| Event       | 8.15     | 8.89 | 8.75     | 7.35 | 3.16        | 3.01| 8.59          | 8.73 |
| Product     | 0.85     | 1.06 | 0.81     | 0.63 | 1.39        | 0.33| 3.99          | 3.99 |
| Total Examples | 106926 | 2031 | 112827  | 2216| 130149      | 11971| 37622         | 11313|

Table 1: Coverage (% of examples) of entity-types in QA datasets. Since examples can mention more than one entity, columns can sum to > 100%. Most datasets except SQuAD are rich in people.

1For NQ, we only consider questions with short answers.
2https://cloud.google.com/natural-language/docs/
3analyzing-entities

We analyze the dev fold, which is consistent with the training fold (Table 1 and 2), as we examine accuracy.

* Work completed while at Google Research
the answer—“who founded Sikhism?” (A: Guru Nanak), in the question—“what did Clara Barton found?” (A: American Red Cross), or the title of the source document—“what play featuring General Uzi premiered in Lagos in 2001?” (A: King Baabu) is in the page on Wole Soyinka. We search until we find an entity: first in the answer, then the question if no entity is found in the answer, and finally the document title.

Demographics are a natural way to categorize these entities and we consider the high-coverage demographic characteristics from Wikidata. Given an entity, Wikidata has good coverage for all datasets: gender (> 99%), nationality (> 93%), and profession (> 94%). For each characteristic, we use the knowledge base to extract the specific value for a person (e.g., the value “poet” for the characteristic “profession”). However, the values defined by Wikidata have inconsistent granularity, so we collapse near-equivalent values (e.g., “writer”, “author”, “poet”, etc. See Appendix A.1–A.2 for an exhaustive list). For questions with multiple values (where multiple entities appear in the answer, or a single entity has multiple values), we create a new value concatenating them together. An ‘others’ value subsumes values with fewer than fifteen examples; people without a value become ‘not found’ for that characteristic.

Three authors manually verify entity assignments by vetting fifty random questions from each dataset. Questions with at least one entity had near-perfect 96% inter-annotator agreement for CLOUD-NL’s annotations, while for questions where CLOUD-NL didn’t find any entity, agreement is 98%. Some errors were benign: incorrect entities sometimes retain correct demographic values; e.g., Elizabeth II instead of Elizabeth I. Other times, coarse-grained nationality ignores nuance, such as the distinction between Greece and Ancient Greece.

### 2.2 Who is in Questions?

Our demographic analysis reveals skews in all datasets, reflecting differences in task focus (Table 2). NQ is sourced from search queries and skews toward popular culture. QB nominally reflects an undergraduate curriculum and captures more “academic” knowledge. TriviaQA is popular trivia, and SQuAD reflects Wikipedia articles.

Across all datasets, men are asked about more than women, and the US is the subject of the majority of questions except in TriviaQA, where the plurality of questions are about the UK. NQ has the highest coverage of women through its focus on entertainment (Film/TV, Music and Sports).

### 3 What Questions can QA Answer?

QA datasets have different representations of demographic characteristics; is this focus benign or do these differences carry through to model accuracy? We analyze a SOTA system for each of our four tasks. For NQ and SQuAD, we use a fine-tuned BERT (Alberti et al., 2019) with curated training data (e.g., downsample questions without answers and split documents into multiple training instances). For the open-domain TriviaQA task, we use ORQA (Lee et al., 2019) that uses BERT-based reader and retriever components. Finally, for QB, we use the competition winner from Wallace et al. (2019), a BERT-based reranker of a TF-IDF retriever.

Accuracy (exact-match) and average F1 are both common QA metrics (Rajpurkar et al., 2016). Since both are related and some statistical tests require binary scores, we focus on exact-match.

Rather than focus on aggregate accuracy, we focus on demographic subsets’ accuracy (Figure 1). For instance, while 66.2% of questions about people are correct in QB, the number is lower for the Dutch (Netherlands) (55.6%) and higher for Ireland (87.5%). Unsurprisingly, accuracy is consistently low on the ‘not_found’ subset, where Wikidata lacks a person’s demographic value.

Are the differences we observe across strata significant? We probe this in two ways: using χ² testing (Plackett, 1983) to see if trends exist and using logistic regression to explore those that do.

| Value | NQ Train | NQ Dev | QB Train | QB Dev | SQuAD Train | SQuAD Dev | TriviaQA Train | TriviaQA Dev |
|-------|----------|--------|----------|--------|-------------|-----------|----------------|--------------|
| Male  | 39.67    | 36.41  | 36.18    | 37.34  | 41.22       | 41.92     | 50.77          | 50.36        |
| Female| 27.47    | 27.56  | 27.09    | 28.87  | 33.44       | 32.50     | 40.54          | 40.29        |
| No Gender | 0.31 | 0.47 | 0.35 | 0.39 | 0.27 | 0.00 | 0.30 | 0.35 |
| US | 39.62 | 38.66 | 29.70 | 28.28 | 32.74 | 24.93 | 31.32 | 30.91 |
| UK | 15.76 | 15.78 | 17.92 | 17.68 | 19.66 | 16.83 | 41.92 | 41.32 |
| France | 1.70 | 1.18 | 10.06 | 7.34 | 7.76 | 10.57 | 4.37 | 4.84 |
| Italy | 1.83 | 1.88 | 8.07 | 10.50 | 9.00 | 3.88 | 3.75 | 3.48 |
| Germany | 1.52 | 2.12 | 7.21 | 6.71 | 4.77 | 6.61 | 3.01 | 3.00 |
| No country | 4.82 | 4.36 | 7.12 | 6.79 | 3.48 | 2.56 | 6.19 | 6.10 |

Table 2: Coverage (% of examples) of demographic values across examples with people in QA datasets. Men dominate, as do Americans.
3.1 Do Demographic Values Affect Accuracy?

The $\chi^2$ test is a non-parametric test of whether two variables are independent. To see if accuracy and characteristics are independent, we apply a $\chi^2$ test to a $n \times 2$ contingency table with $n$ rows representing the frequency of that characteristic’s subsets contingent on whether the model prediction is correct or not (Table 3). If we reject the null with a Bonferroni correction (Holm, 1979, divide the $p$-value threshold by three, as we have multiple tests for each dataset), that suggests possible relationships: gender in NQ ($p = 2.36 \times 10^{-12}$), and professional field in NQ ($p = 0.0142$), QB ($p = 2.34 \times 10^{-7}$) and TriviaQA ($p = 0.0092$). However, we find no significant relationship between nationality and accuracy in any dataset.

While $\chi^2$ identifies which characteristics impact model accuracy, it does not characterize how. For instance, $\chi^2$ indicates NQ’s gender is significant, but is this because accuracy is higher for women, or because the presence of both genders in examples lowers the accuracy?

3.2 Exploration with Logistic Regression

Thus, we formulate a simple logistic regression: can an example’s demographic values predict if a model answers correctly? Logistic regression and related models are the workhorse for discovering and explaining the relationship between variables in history (McCloskey and McCloskey, 1987), education (van der Linden and Hambleton, 2013), political science (Poole and Rosenthal, 2011), and sports (Glickman and Jones, 1999). Logistic regression is also a common tool in NLP: to find linguistic constructs that allow determiner omission (Kiss et al., 2010) or to understand how a scientific paper’s attributes effect citations (Yogatama et al., 2011). Unlike model calibration (Niculescu-Mizil and Caruana, 2005), whose goal it to maximize prediction accuracy, the goal here is explanation.

We define binary features for demographic values which $\chi^2$ test found significant (thus, SQaud, the nationality characteristic, and gender characteristic for all but NQ are excluded). For instance, a question about Abidali Neemuchwala would have features for g_male, o_executive but zero for everything else. Real-valued features, multi_entities and multi_answers, capture the effect of multiple
person-entities and multiple gold-answers (scaled with the base two logarithm).

But that is not the only reason an answer may be difficult or easy. Following Sugawara et al. (2018), we incorporate features that reveal the questions’ difficulty. For instance, questions that clearly hint the answer type reduce ambiguity. The \( t_{\text{who}} \) checks if the token “who” is in the start of the question. Similarly, \( t_{\text{what}}, t_{\text{when}} \), and \( t_{\text{where}} \) capture other entity-types. Questions are also easier if evidence only differs from the question by a couple of words; thus, \( q_{\text{sim}} \) is the Jaccard similarity between question and evidence tokens. Finally, the binary feature \( e_{\text{train count}} \) marks if the person-entities occur in training data more than twice.

We first drop features with negligible effect on accuracy using LASSO (regularization \( \lambda = 1 \)) by removing zero coefficients. For remaining features, Wald statistics (Fahrmeir et al., 2007) estimate \( p \)-values. Although we initially use quadratic features they are all eliminated during feature reduction. Thus, we only report the linear features with a minimal significance (\( p \)-value < 0.1).

### 3.3 How do Properties Affect Accuracy?

Recall that logistic regression uses features to predict whether the QA system will get the answer right or not. Features associated with correct answers have positive weights (like those derived from Sugawara et al. (2018), \( q_{\text{sim}} \) and \( e_{\text{train count}} \)), those associated with incorrect answers have negative weights, and features without effect will be near zero. Among the \( t_{\text{who}} \) features, \( t_{\text{who}} \) significantly correlates with model correctness, especially in NQ and QB, where questions asked directly about a person.

However, our goal is to see if, after accounting for obvious reasons a question could be easy, demographic properties can explain QA accuracy. The strongest effect is for professions (Table 4). For instance, while NQ and QB systems struggle on science questions, TriviaQA’s does not. Science has roughly equivalent representation (Table 2), suggesting QB questions are harder.

While \texttt{multi_answer} (and \texttt{multi_entities}) reveal harder NQ questions, it has a positive effect in TriviaQA, as TriviaQA uses multiple answers for alternate formulations of answers (Appendix B.2.1, B.2.2), which aids machine reading, while multiple NQ answers are often a sign of ambiguity (Boyd-Graber and Börschinger, 2020; Si et al., 2021): “Who says that which we call a rose?” A: Juliet, A: William Shakespeare. For male and female genders, NQ has no statistically significant effect on accuracy, only questions about entities with multiple genders depresses accuracy. Given the many findings of gender bias in NLU (Zhao et al., 2017; Webster et al., 2018; Zhao et al., 2018; Stanovsky et al., 2019), this is surprising. However, we caution against accepting this conclusion without further investigation given the strong correlation of gender with professional field (Goulden et al., 2011), where we do see significant effects.

Taken together, the \( \chi^2 \) and logistic regression analysis give us reason to be optimistic: although data are skewed for all subsets, QA systems might well generalize from limited training data across gender and nationality.

### 4 Related Work

Language is a reflection of culture. Like other cultural artifacts—encyclopedias (Reagle and Rhue, 2011), and films (Sap et al., 2017)—QA has more men than women. Other artifacts like children’s books have more gender balance but reflect other aspects of culture (Larrick, 1965).

The NLP literature is also grappling with demographic discrepancies. Standard coreference systems falter on gender-balanced corpora (Webster et al., 2018), and Zhao et al. (2018) create synthetic training data to reduce bias. Similar coreference issues plague machine translation systems (Stanovsky et al., 2019), and Li et al. (2020) use QA to probe biases of NLP systems. Sen and Saffari (2020) show that there are shortcomings in QA datasets and evaluations by analysing their out-of-domain generalization capabilities and ability to handle question variation. Joint models of vision and language suggest that biases come from language, rather than from vision (Ross et al., 2021). However, despite a range of mitigation techniques (Zhao et al., 2017; inter alia) none, to our knowledge, have been successfully applied to QA, especially from the demographic viewpoint.

### 5 Discussion and Conclusion

This paper delivers both good news and bad news. While datasets remain imperfect and reflect societal imperfections, for many demographic properties, we do not find strong evidence that QA suffers from this skew.

However, this is an absence of evidence rather
than evidence of absence: these are skewed datasets that have fewer than a quarter of the questions about women. It is difficult to make confident assessments on such small datasets—many demographic values were excluded because they appeared infrequently (or not at all). Improving the diversity of QA datasets can help us be more certain that QA systems do generalize and reflect the diverse human experience. Considering such shortcomings, Rodriguez et al. (2021) advocate improving evaluation by focusing on more important examples for ranking models; demographic properties could further refine more holistic evaluations.

A broader analysis beyond person entities would indeed be a natural extension of this work. Label propagation can expand the analysis beyond people: the Hershey-Chase experiment is associated with Alfred Hershey and Martha Chase, so it would—given the neighboring entities in the Wikipedia link graph—be 100% American, 50% male, and 50% female. Another direction for future work is accuracy under counterfactual perturbation: swapping real-world entities (in contrast with nonce entities in Li et al. (2020)) with different demographic values.

Nonetheless, particularly for professional fields, imbalances remain. The lack of representation in QA could cause us to think that things are better than they are because of Simpson’s paradox (Blyth, 1972); gender and profession are not independent! For example, in NQ, our accuracy on women is higher in part because of its tilt toward entertainment, and we cannot say much about women scientists. We therefore caution against interpreting strong model performance on existing QA datasets as evidence that the task is ‘solved’. Instead, future work must consider better dataset construction strategies and robustness of accuracy metrics to different subsets of available data, as well as unseen examples.

**Acknowledgements**

We thank Michael Collins, Slav Petrov, Tulsee Doshi, Sephora Madjiheurem, Benjamin Börschinger, Pedro Rodriguez, Massimiliano Ciaramita, Kenton Lee, Alex Beutal, Kenton Lee, and Emily Pitler for their early and insightful comments on the proposal and drafts. Additionally, insights about Google’s CLOUD-NL entity linking tool and WIKIDATA KB from Jan Botha, Denny Vrandecic, Livio Soares and Tom Kwiatkowski were quite useful in designing the entity linking and attribute extraction pipeline.

---

**Table 4:** Influential features after filtering characteristics based on a $\chi^2$ test (Figure 1). Highly influential features ($p$-value < 0.1), both positive (blue) and negative (red). Higher number of *’s signals higher significance.

| Dataset | Features          | Coef | SE     | Wald ($|W|$) | $P_{Z \sim N(|Z| > |W|)}$ |
|---------|-------------------|------|--------|-----------|--------------------------|
| NQ      | bias              | 1.964| 0.727  | 2.703     | 0.0069  | ***                      |
|         | multi_answers     | -1.893| 0.438  | 4.327     | 0.0000  | ****                    |
|         | o_not_found       | -1.112| 0.514  | 2.163     | 0.0305  | **                      |
|         | o_science/tech    | +0.773| 0.280  | 2.764     | 0.0057  | ***                      |
|         | multi_entities    | -0.678| 0.342  | 1.979     | 0.0478  | **                      |
|         | q_sim             | +0.406| 0.210  | 1.934     | 0.0531  | *                       |
|         | e_train_count     | +0.353| 0.178  | 1.979     | 0.0479  | **                      |
| Model Fit: 78.09% |                  |      |        |           |                  |
| QB      | e_train_count     | +1.922| 0.269  | 7.144     | 0.0000  | ****                    |
|         | bias              | -1.024| 0.291  | 3.516     | 0.0004  | ****                    |
|         | o_film/tv         | -0.910| 0.470  | 1.934     | 0.0531  | *                       |
|         | multi_entities    | -0.870| 0.165  | 5.287     | 0.0000  | ****                    |
|         | o_science/tech    | -0.667| 0.362  | 1.812     | 0.0700  | *                       |
|         | o_religion        | -0.655| 0.562  | 1.812     | 0.0700  | *                       |
|         | o_writing         | +0.402| 0.189  | 2.128     | 0.0334  | **                      |
|         | t_who             | +0.363| 0.129  | 1.777     | 0.0754  | *                       |
| Model Fit: 71.90% |                  |      |        |           |                  |
| TriviaQA| bias              | -1.066| 0.114  | 9.353     | 0.0000  | ****                    |
|         | o_religion        | -0.443| 0.255  | 1.738     | 0.0822  | *                       |
|         | o_law/crime       | +0.412| 0.218  | 1.890     | 0.0588  | *                       |
|         | multi_answers     | +0.341| 0.024  | 14.090    | 0.0000  | ****                    |
|         | t_who             | +0.230| 0.129  | 1.778     | 0.0754  | *                       |
|         | o_politics        | -0.208| 0.095  | 2.177     | 0.0295  | **                      |
|         | o_writing         | +0.192| 0.098  | 1.955     | 0.0506  | *                       |
| Model Fit: 60.18% |                  |      |        |           |                  |
Ethical Considerations

This work analyses demographic subsets across QA datasets based on Gender, Nationality and Profession. We believe the work makes a positive contribution to representation and diversity by pointing out the skewed distribution of existing QA datasets. To avoid noise being interpreted as signal given the lack of diversity in these datasets, we could not include various subgroups that we believe should have been part of this study: non-binary, intersectional groups (e.g., women scientists in NQ), people indigenous to subnational regions, etc. We believe increasing representation of all such groups in QA datasets would improve upon the status quo. We infer properties of mentions using Google Cloud-NL to link the entity in a QA example to an entry in the WIKIDATA knowledge base to attribute gender, profession and nationality. We acknowledge that this is not foolproof and itself vulnerable to bias, although our small-scale accuracy evaluation did not reveal any concerning patterns.

All human annotations are provided by authors to verify entity-linkings and were fairly compensated.

References

Chris Alberti, Kenton Lee, and Michael Collins. 2019. A BERT baseline for the natural questions. arXiv preprint arXiv:1901.08634.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. CoRR, abs/1910.11856.

Colin R. Blyth. 1972. On Simpson’s paradox and the sure-thing principle. Journal of the American Statistical Association, 67(338):364–366.

Jordan Boyd-Graber and Benjamin Börschinger. 2020. What question answering can learn from trivia nerds. In Proceedings of the Association for Computational Linguistics.

Jordan Boyd-Graber, Brianna Satinoff, He He, and Hal Daume III. 2012. Besting the quiz master: Crowdsourcing incremental classification games. In Proceedings of Empirical Methods in Natural Language Processing.

Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian Marx. 2007. Regression. Springer.

David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Chris Welty. 2010. Building Watson: An Overview of the DeepQA Project. AI Magazine, 31(3).

Mark E Glickman and Albyn C Jones. 1999. Rating the chess rating system. Chance, 12.

Marc Goulden, Mary Ann Mason, and Karie Frasch. 2011. Keeping women in the science pipeline. The Annals of the American Academy of Political and Social Science, 638:141–162.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6(2):65–70.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Proceedings of the Association for Computational Linguistics.

Tibor Kiss, Katja Keßelmeier, Antje Müller, Claudia Roch, Tobias Stadtfeld, and Jan Strunk. 2010. A logistic regression model of determiner omission in PPs. In Coling 2010: Posters, pages 561–569, Beijing, China. Coling 2010 Organizing Committee.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polousukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Liion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. Transactions of the Association for Computational Linguistics.

Nancy Larrick, 1965. The all-white world of children’s books. The Saturday Review, 64:63–65.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open-domain question answering. In Proceedings of the Association for Computational Linguistics.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7315–7330, Online. Association for Computational Linguistics.

Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. 2020. UNQOVERing stereotyping biases via underspecified questions. In Findings of the Association for Computational Linguistics: EMNLP.

Deirdre N. McCloskey and Donald N. McCloskey. 1987. Econometric History. Casebook Series. Macmillan Education.

Sewon Min, Julian Michael, Hananah Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 5783–5797, Online. Association for Computational Linguistics.
Adam Najberg. 2018. Alibaba AI model tops humans in reading comprehension.

Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In Proceedings of the International Conference of Machine Learning.

Robin L. Plackett. 1983. Karl Pearson and the chi-squared test. International Statistical Review/Revue Internationale de Statistique, pages 59–72.

K.T. Poole and H.L. Rosenthal. 2011. Ideology and Congress. American Studies. Transaction Publishers.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In Proceedings of Empirical Methods in Natural Language Processing.

Joseph Reagle and Lauren Rhue. 2011. Gender bias in Wikipedia and Britannica. International Journal of Communication, 5(0).

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change nlp leaderboards? Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).

Candace Ross, Boris Katz, and Andrei Barbu. 2021. Measuring social biases in grounded vision and language embeddings. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 998–1008, Online. Association for Computational Linguistics.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In Proceedings of Empirical Methods in Natural Language Processing.

Priyanka Sen and Amir Saffari. 2020. What do models learn from question answering datasets? Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Chenglei Si, Chen Zhao, and Jordan Boyd-Graber. 2021. What’s in a name? answer equivalence for open-domain question answering. In Empirical Methods in Natural Language Processing.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In Proceedings of the Association for Computational Linguistics.

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. What makes reading comprehension questions easier? In Proceedings of Empirical Methods in Natural Language Processing.

Wim J van der Linden and Ronald K Hambleton. 2013. Handbook of modern item response theory. Springer Science & Business Media.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019. Trick me if you can: Human-in-the-loop generation of adversarial question answering examples. Transactions of the Association of Computational Linguistics, 10.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. Transactions of the Association for Computational Linguistics, 6:605–617.

Dani Yogatama, Michael Heilman, Brendan O’Connor, Chris Dyer, Bryan R. Routledge, and Noah A. Smith. 2011. Predicting a scientific community’s response to an article. In Proceedings of Empirical Methods in Natural Language Processing.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In Proceedings of Empirical Methods in Natural Language Processing.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In Conference of the North American Chapter of the Association for Computational Linguistics.
Appendix

A Entity collapses of demographic values

While mapping QA examples to person entities and values for their corresponding demographic characteristics (Section 2), we encountered many nearby values: ‘Poet’, ‘Writer’, ‘Author’. We collapse such values into a single label which we use for further analysis. This section enlists all the collapses that we encounter for determining nationality of people (Appendix A.1) and their professions (Appendix A.2).

A.1 Entity-collapses for Nationality values

**US**: kingdom of hawaii, united states, united states of america

**UK**: commonwealth of england, great britain, kingdom of england, kingdom of mercia, kingdom of scotland, kingdom of wessex, united kingdom, united kingdom of great britain and ireland

**Albania**: kingdom of albania

**Austria**: austrian empire, federal state of austria, first republic of austria

**Cyprus**: kingdom of cyprus, republic of cyprus, turkish republic of northern cyprus

**Denmark**: kingdom of denmark

**France**: kingdom of france

**Germany**: german confederation, german democratic republic, german empire, german reich, germany, kingdom of hanover, kingdom of prussia, kingdom of saxony, nazi germany, north german confederation, prussia, republic of german-austria, west germany

**Greece**: ancient greece, greece

**Hungary**: hungary, kingdom of hungary, people’s republic of hungary

**Ireland**: irish republic, kingdom of ireland

**Italy**: ancient rome, florence, holy roman empire, kingdom of italy, kingdom of sardinia

**Netherlands**: dutch republic, kingdom of the netherlands

**Poland**: kingdom of poland, poland

**Portugal**: kingdom of portugal

**Romania**: kingdom of romania, romania, socialist republic of romania

**Spain**: crown of castile, kingdom of aragon, kingdom of castile, kingdom of navarre, spain

**Yugoslavia**: federal republic of yugoslavia, kingdom of yugoslavia, socialist federal republic of yugoslavia, yugoslavia

**Iraq**: ba’athist iraq, iraq, kingdom of iraq, mandatory iraq, republic of iraq (1958–68)

**Israel**: israel, kingdom of israel, land of israel

**Russia**: russia, russian empire, russian soviet federative socialist republic, soviet union, tsardom of russia

**India**: british raj, delhi sultanate, dominion of india, india
China: china, people’s republic of china, republic of china (1912-1949)

Egypt: ancient egypt, egypt, kingdom of egypt, republic of egypt

A.2 Entity-collapses for Profession values

Writing: author, biographer, cartoonist, children’s writer, comedy writer, comics artist, comics writer, contributing editor, cookery writer, detective writer, diarist, editor, editorial columnist, essayist, fairy tales writer, grammarian, hymnwriter, journalist, lexicographer, librettist, linguist, literary, literary critic, literary editor, literary scholar, memoirist, newspaper editor, non-fiction writer, novelist, opinion journalist, philologist, photojournalist, physician writer, playwright, poet, poet lawyer, preface author, prosaist, religious writer, science fiction writer, science writer, scientific editor, screenwriter, short story writer, tragedy writer, travel writer, women letter writer, writer

Sports: amateur wrestler, american football coach, american football player, archer, artistic gymnast, association football manager, association football player, association football referee, athlete, athletics competitor, australian rules football player, badminton player, ballet dancer, ballet master, ballet pedagogue, baseball player, basketball coach, basketball player, biathlete, biathlon coach, boxer, bridge player, canadian football player, chess player, choreographer, coach, cricket umpire, cricketer, dancer, darts player, field hockey player, figure skater, figure skating choreographer, figure skating coach, formula one driver, gaelic football player, golfer, gridiron football player, gymnast, head coach, hurler, ice dancer, ice hockey coach, ice hockey player, jockey, judoka, lacrosse player, long-distance runner, marathon runner, marimba player, martial artist, middle-distance runner, mixed martial artist, motorcycle racer, poker player, polo player, pool player, professional wrestler, quidditch player, racing automobile driver, racing driver, rink hockey player, rugby league player, rugby player, rugby union coach, rugby union player, runner, short track speed skater, skateboarder, skeleton racer, snooker player, snowboarder, sport cyclist, sport shooter, sporting director, sports agent, sports commentator, sprinter, squash player, surfer, swimmer, table tennis player, taekwondo athlete, tennis coach, tennis player, thai boxer, track and field coach, viol player, volleyball player, water polo player

Music: bass guitar, bassist, blues musician, child singer, classical composer, classical guitarist, classical pianist, collector of folk music, composer, conductor, country musician, drummer, film score composer, ghost singer, guitar maker, guitarist, heavy metal singer, instrument maker, instrumentalist, jazz guitarist, jazz musician, jazz singer, keyboardist, lyricist, multi-instrumentalist, music arranger, music artist, music critic, music director, music interpreter, music pedagogue, music pedagogy, music producer, music publisher, music theorist, music video director, musical, musical instrument maker, musician, musicologist, opera composer, opera singer, optical instrument maker, organist, pianist, playback singer, professor of music composition, rapper, record producer, recording artist, rock drummer, rock musician, saxophonist, session musician, singer, singer-songwriter, songwriter, violinist

Fictional: fictional aviator, fictional businessperson, fictional character, fictional cowboy, fictional domestic worker, fictional firefighter, fictional journalist, fictional mass murderer, fictional pirate, fictional police officer, fictional politician, fictional schoolteacher, fictional scientist, fictional seaman, fictional
Politics: activist, ambassador, animal rights advocate, anti-vaccine activist, civil rights advocate, civil servant, climate activist, colonial administrator, consort, dictator, diplomat, drag queen, duke, emperor, feminist, foreign minister, government agent, governor, human rights activist, internet activist, khan, king, leader, lgbt rights activist, military commander, military leader, military officer, military personnel, military theorist, minister, monarch, peace activist, political activist, political philosopher, political scientist, political theorist, politician, president, prince, princess, protestant reformer, queen, queen consort, queen regnant, religious leader, revolutionary, ruler, secretary, social reformer, socialite, tribal chief

Artist: architect, artist, baker, blacksmith, car designer, chef, costume designer, design, designer, fashion designer, fashion photographer, fresco painter, furniture designer, game designer, glass artist, goldsmith, graffiti artist, graphic artist, graphic designer, house painter, illustrator, industrial designer, interior designer, jewellery designer, landscape architect, landscape painter, lighting designer, painter, photographer, postage stamp designer, printmaker, production designer, scientific illustrator, sculptor, sound designer, textile designer, type designer, typographer, visual artist

Film/tv: actor, character actor, child actor, documentary filmmaker, dub actor, factory owner, fashion model, film actor, film critic, film director, film editor, film producer, filmmaker, glamour model, line producer, model, pornographic actor, reality television participant, runway model, television actor, television director, television editor, television presenter, television producer, voice actor

Executive: bank manager, business executive, business magnate, businessperson, chief executive officer, entrepreneur, executive officer, executive producer, manager, real estate entrepreneur, talent manager

Stage: circus performer, comedian, entertainer, mime artist, musical theatre actor, stage actor, stand-up comedian, theater director

Law/crime: art thief, attorney at law, bank robber, canon law jurist, courtier, criminal, judge, jurist, lawyer, official, private investigator, robber, serial killer, spy, thief, war criminal

History: anthropologist, archaeologist, art historian, church historian, classical archaeologist, egyptologist, explorer, historian, historian of classical antiquity, historian of mathematics, historian of science, historian of the modern age, labor historian, legal historian, literary historian, military historian, music historian, paleoanthropologist, paleontologist, philosophy historian, polar explorer, scientific explorer

Science/tech: aerospace engineer, alchemist, anesthesiologist, artificial intelligence researcher, astrologer, astronaut, astronomer, astrophysicist, auto mechanic, bacteriologist, biochemist, biologist, botanist, bryologist, cardiologist, chemical engineer, chemist, chief engineer, civil engineer, climatologist, cognitive scientist, combat engineer, computer scientist, cosmologist, crystallographer, earth scientist, ecologist, educational psychologist, electrical engineer, engineer, environmental scientist, epidemiologist,
ethnologist, ethologist, evolutionary biologist, geochemist, geographer, geologist, geophysicist, immunologist, industrial engineer, inventor, marine biologist, mathematician, mechanic, mechanical automaton engineer, mechanical engineer, meteorologist, microbiologist, mining engineer, naturalist, neurologist, neuroscientist, nuclear physicist, nurse, ontologist, ornithologist, patent inventor, pharmacologist, physician, physicist, physiologist, planetary scientist, psychiatrist, psychoanalyst, psychologist, railroad engineer, railway engineer, research assistant, researcher, scientist, social psychologist, social scientist, sociologist, software engineer, space scientist, statistician, structural engineer, theoretical biologist, theoretical physicist, virologist, zoologist

**Polymath:** polymath

**Education:** academic, adjunct professor, associate professor, educator, head teacher, high school teacher, history teacher, lady margaret’s professor of divinity, pedagogue, professor, school teacher, sex educator, teacher, university teacher

**Economics:** economist

**Religion:** anglican priest, bible translator, bishop, catholic priest, christian monk, lay theologian, monk, pastor, pope, preacher, priest, theologian

**Military:** air force officer, aircraft pilot, commanding officer, fighter pilot, general officer, helicopter pilot, intelligence officer, naval officer, officer of the french navy, police officer, soldier, starship pilot, test pilot

**Translation:** translator

**Philosophy:** analytic philosopher, philosopher, philosopher of language,
B Logistic Regression features.

This section enlists a full set of features used for the logistic regression analysis after feature reduction, each with their coefficients, standard error, Wald Statistic and significance level in Table 5. We also describe the templates and the implementation details of the features using in our logistic regression analysis (Section 3.2) in Appendix B.1, and finally enlist some randomly sampled examples both from NQ and TriviaQA datasets in Appendix B.2 to show how multi_answers feature has disparate effects on them.

B.1 Implementation of Logistic Regression features

- **q_sim**: For closed-domain QA tasks like NQ and SQuAD, this feature measures (sim)ilarity between (q)uestion text and evidence sentence—the sentence from the evidence passage which contains the answer text—using Jaccard similarity over unigram tokens (Sugawara et al., 2018). Since we do not include SQuAD in our logistic regression analysis (Section 3.2), this feature is only relevant for NQ.

- **e_train_count**: This binary feature represents if distinct (e)ntities appearing in a QA example (through the approach described in Section 2) appears more than twice in the particular dataset’s training fold. We avoid logarithm here as even the log frequency for some commonly occurring entities exceeds the expected feature value range.

- **t_wh***: This represents the features that captures the expected entity type of the answer: t_who, t_what, t_where, t_when. Each binary feature captures if the particular “wh*” word appears in the first ten (t)okens of the question text.  

- **multi_entities**: For number of linked person-entities in a example as described in Section 2 as n, this feature is \( \log_2(n) \). Hence, this feature is 0 for example with just one person entity.

- **multi_answers**: For number of gold-answers annotated in a example as n, this feature is \( \log_2(n) \). Hence, this feature is 0 for example with just one answer.

- **g_***: Binary demographic feature signaling the presence of the (g)ender characterized by the feature. For instance, g_female signals if the question is about a female person.

- **o_***: Binary demographic feature signaling the presence of the occupation (or profession) as characterized by the feature. For instance, o_writer signals if the question is about a writer.

B.2 Examples with multi_answers feature

In the Logistic Regression analysis (Section 3.2), we create two features: multi_answers and multi_entities. Former captures the presence of multiple gold answers to the question in a given example, while latter signals presence of multiple person entities — all in either the answers, the question text or the document title for a given example. While multi_entities has consistent negative co-relation with model correctness (Appendix B), multi_answers has a disparate effect. Though it signals towards incorrectly answered examples in NQ, it has a statistically significant positive correlation with model correctness for TriviaQA examples. Going through the examples, it reveals that TriviaQA uses multiple answers to give alternate formulations of an answer, which aids machine reading, while multiple NQ answers are often a sign of question ambiguity (Min et al., 2020).

To demonstrate that, we enlist here examples from development fold of both NQ (Appendix B.2.1) and TriviaQA (Appendix B.2.2) that have multiple gold answers.

B.2.1 NQ examples with multiple answers:

| id          | Question                                                                                     | QA                           |
|-------------|----------------------------------------------------------------------------------------------|------------------------------|
| 413520984491483842 | Q: who carried the us flag in the 2014 olympics                                             | A: Todd Lodwick               |
| 82338716539218945006 | Q: who says that which we call a rose                                                       | A: William Shakespeare       |
| 3280415545011911929  | Q: who has won the most superbowls as a player                                              | A: Charles Haley             |
| 3248410603422198181  | Q: who has written the song if i were a boy                                                  | A: BC Jitri                   |
| 3248410603422198181  | Q: who started the guinness book of world records                                           | A: Jack McWhirter            |
| 3248410603422198181  | Q: who was the nurse on andy griffith show                                                  | A: Linda                     |
| 3248410603422198181  | Q: who wrote the song if i were a boy                                                        | A: Jack McWhirter            |
| 3248410603422198181  | Q: who conducted the opening concert at carnegie hall                                       | A: Todd McWhirter            |

QB questions often start with “For 10 points, name this writer who...”
Table 5: Influential features revealed through Logistic Regression Analysis (Sec 3.2) over the demographic characteristics deemed significant through the $\chi^2$ test (Figure 1). We report the highly influential features with significance of $p < 0.1$, both positive (blue) and negative (red), and bold the highly significant ones ($p < 0.05$). Number of * in the last column represents the significance level of that feature.

| Dataset | Features | Coef  | SE   | Wald (W) | $p = \chi^2 (|Z| > |W|) |  |
|---------|----------|-------|------|----------|-------------------------|---|
|.bias   | bias     | -1.066| 0.114| 9.353    | 0.000                   | ⋆
|        | education| -0.665| 0.554| 1.245    | 0.262                   | ⋆
|        | o_economics| -0.503| 0.579| 0.869    | 0.354                   | ⋆
|        | o_philosophy| -0.469| 0.342| 1.370    | 0.243                   | ⋆
|        | o_religion| -0.443| 0.255| 1.738    | 0.189                   | ⋆
|        | o_law/crime| -0.442| 0.218| 1.898    | 0.058                   | ⋆
|        | multi_answers| -0.341| 0.024| 14.099   | 0.000                   | ⋆
|        | t_who     | -0.238| 0.129| 1.778    | 0.185                   | ⋆
|        | o_taste   | -0.212| 0.257| 0.824    | 0.365                   | ⋆
|        | o_politics| -0.208| 0.095| 2.177    | 0.032                   | ⋆
|        | o_movie   | -0.183| 0.117| 1.570    | 0.214                   | ⋆
|        | o_artist  | -0.150| 0.092| 1.626    | 0.101                   | ⋆
|        | o_politics| -0.146| 0.099| 0.408    | 0.525                   | ⋆
|        | o_executive| -0.143| 0.068| 0.699    | 0.405                   | ⋆
|        | o_others  | -0.101| 0.052| 0.665    | 0.418                   | ⋆
|        | multi_answers| -0.099| 0.077| 1.290    | 0.209                   | ⋆
|        | e_train_count| -0.091| 0.075| 1.079    | 0.296                   | ⋆
|        | t_who     | -0.018| 0.134| 0.132    | 0.294                   | ⋆

A: Walter Donovan  
A: Peter Stack Esplinovsky  
id: 3773753259703193836  
Q: who founded amazon where is the headquarters of amazon  
A: founded by Jeff Bezos  
A: Based in Seattle, Washington  

Table 4: Significant answers found through subject generation (Sec 3.1) that are notable due to their frequency (.37) or are highly significant ($p < 0.05$) through the $\chi^2$ test (Figure 1). The counts are the number of times each answer was found in the generated questions of the participants.

A: George Harrison  
id: 567427281639928690  
Q: who sings you're welcome in moana credits  
A: Lin-Manuel Miranda  
A: Jordan Fisher  
id: 743232250575146771  
Q: who wrote the song i hate you i love you  
A: Garrett Nosh  
A: Omega lyrics  
id: 753279777079537148  
Q: who is the owner of reading football club  
A: Xin Li Dai  
A: Younger Fly  
id: 7136132803738849961  
Q: who placed guitar on my guitar gently weeps  
A: Eric Clapton  
A: George Harrison  
id: 731603818131321387  
Q: who sang the theme song to that 70s show  
A: Todd Griffin  
A: Young fly  
id: 733553410879363348  
Q: who came up with the initial concept of protons and neutrons  
A: Werner Heisenberg  
A: Future Foundation  
id: 73160429334516617  
Q: who missed the plane the day the music died  
A: Warren Jennings  
A: Leon Russell  

A: Ringo Starr  
A: George Harrison  
id: 567427281639928690  
Q: who sang the song you're welcome in moana credits  
A: Lin-Manuel Miranda  
A: Jordan Fisher  
id: 743232250575146771  
Q: who wrote the song i hate you i love you  
A: Garrett Nosh  
A: Omega lyrics  
id: 753279777079537148  
Q: who is the owner of reading football club  
A: Xin Li Dai  
A: Younger Fly  
id: 7136132803738849961  
Q: who placed guitar on my guitar gently weeps  
A: Eric Clapton  
A: George Harrison  
id: 731603818131321387  
Q: who sang the theme song to that 70s show  
A: Todd Griffin  
A: Young fly  
id: 733553410879363348  
Q: who came up with the initial concept of protons and neutrons  
A: Werner Heisenberg  
A: Future Foundation  
id: 73160429334516617  
Q: who missed the plane the day the music died  
A: Warren Jennings  
A: Leon Russell  

A: Ringo Starr  

Q: What is Robin Williams character called in Good Morning Vietnam?
A: Adrian

Q: Who was the first name of the jazz trombonist Kid Ory?
A: Edward

Q: Which of Queen Elizabeth’s children is the lowest in succession to (i.e. furthest away from) the throne?
A: Anne

Q: “Which radio comedian’s catchphrase was “isn’t daft as a brush”?”
A: KEN PLATT

Q: According to Sammy Hagger, what can’t he drive?
A: A
dr

drive

Q: What was Grace Darling’s father’s job?
A: LIGHTHOUSE KEEPER

Q: What was the name of the older brother of Henry 8th?
A: Arthur

Q: A man was killed in an explosion in his pint because his name was one of the ingredients?
A: Henry 8th

Q: In which year did both T-Rex’s Marc Bolan and Elvis Presley die?
A: 1977

Q: Who played Hotlips Houlahan in the 1972 film MASH?
A: Sally Kellerman

Q: Who was the middle of the author William Thackeray’s “Dialogue” off their list of banned books, and Britain repeal the death penalty for over 100 crimes?
A: five

Q: Who wrote the novel ‘The Beach’ on which the film was based?
A: Michael Crichton

Q: What is the title of the most famous painting by Franz Hals?
A: "Portrait of Anne, Countess of Derby"

Q: What was the name of the private eye played by Trevor Eve on TV in the ‘70s?
A: "Terrybundle" (disambiguation)

Q: What was the name of the private eye played by Trevor Eve on TV in the ‘70s?
A: "Terrybundle" (disambiguation)

Q: What was the name of the jazz trombonist Kid Ory?
A: Edward

Q: Which of the Great Train Robbers became a florist outside Waterloo station until he was found hanged in a lock up?
A: Buster Edwards

Q: Who presents the BBC quiz show ‘Perfection’?
A: ALLAN KLEIN

Q: Which Irish poet wrote ‘The Hymn of Jesus’?
A: W.B. Yeats

Q: What was the name of the author William Thackeray’s “Dialogue” off their list of banned books, and Britain repeal the death penalty for over 100 crimes?
A: five

Q: Who was the first correct translation of Egyptian hieroglyphs from the Rosetta Stone, the Roman Catholic Church take Galileo Galilei’s "Dialogue" off their list of banned books, and Britain repeal the death penalty for over 100 crimes?
A: Jean-Francois Champollion

Q: Which of the Great Train Robbers became a florist outside Waterloo station until he was found hanged in a lock up?
A: Buster Edwards

Q: What is the name of the author William Thackeray’s “Dialogue” off their list of banned books, and Britain repeal the death penalty for over 100 crimes?
A: Jean-Francois Champollion
Q: Who wrote the 1951 novel ‘The Caine Mutiny’?
A: HERMAN WOUK

Q: Said to refer erroneously to the temperature at which book paper catches fire, the title of Ray Bradbury’s 1953 novel about a futuristic society in which reading books is illegal, is called ‘Fahrenheit...’ what? 972; 451; 100; or 25?
A: 451

Q: Who was the driver of the limousine at the time of Diana Princess of Wales’ death?
A: Henri Paul

Q: Which island in the Grenadines of St. Vincent was bought by Colin Tennant in 1958? Princess Margaret built a holiday home there in the 1960’s.
A: Mustique

Q: Which pop star had the real name of Ernest Evans?
A: ‘Chubby Checker’

Q: Which supermodel said, “I look very scary in the mornings.”
A: We don’t wake up for less than $10,000 a day
A: Linda Evangelista