Maharshi Gor, Kellie Webster, and **Jordan Boyd-Graber**. **Toward Deconfounding the Influence of Subject's Demographic Characteristics in Question Answering**. *Empirical Methods in Natural Language Processing*, 2021, 9 pages.


```
    @inproceedings{Gor:Webster:Boyd-Graber-2021,
Title = {Toward Deconfounding the Influence of Subject's Demographic Characteristics in Question Answering
Author = {Maharshi Gor and Kellie Webster and Jordan Boyd-Graber},
Booktitle = {Empirical Methods in Natural Language Processing},
Year = {2021},
Location = {Punta Cana},
Pages = {6},
Url = {http://cs.umd.edu/~jbg//docs/2021_emnlp_qa_fairness.pdf},
}
```


Accessible Abstract: The data used to train computer question answering systems have three times as many men as women. This paper examines whether this is a problem for question answering accuracy. After a thorough investigation, we do not find evidence of serious accuracy discrepancies between languages. However, an absence of evidence is not evidence of absence, and we would argue that we need more diverse datasets to better represent the world's population.

Links:

- Research Talk [https://youtu.be/Hopd3oHfoYk]

Contact Jordan Boyd-Graber (jbg@boydgraber.org) for questions about this paper.

Toward Deconfounding the Influence of Entity Demographics for Question Answering Accuracy

Maharshi Gor[*]
CS
University of Maryland
mgor@cs.umd.edu

Kellie Webster
Google Research, New York
websterk@google.com

Jordan Boyd-Graber[*]
CS, UMIACS, iSchool, LCS
University of Maryland
jbg@umiacs.umd.edu

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.

Entity Type	NQ		QB		SQuAD		TriviaQA	
	Train	Dev	Train	Dev	Train	Dev	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	44.32	44.10
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	14.61
Organization	17.73	18.05	20.37	17.78	21.34	19.14	22.15	21.14
Event	8.15	8.89	8.87	8.75	3.16	3.01	8.59	8.73
Product	0.85	1.06	0.88	0.63	1.39	0.33	3.99	3.99
Total Examples	*106926*	*2631*	*112927*	*2216*	*130319*	*11873*	*87622*	*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.

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-NL[2] 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 models 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 the answer—"who founded Sikhism?" (A: *Guru*

[*] Work completed while at Google Research

[1] For NQ, we only consider questions with short answers.

[2] https://cloud.google.com/natural-language/docs/analyzing-entities

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

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.[4] 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 are search queries and skew 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*).

[4] https://www.wikidata.org/wiki/Wikidata:Database_download

	Value	NQ		QB		SQuAD		TriviaQA	
		Train	Dev	Train	Dev	Train	Dev	Train	Dev
Gender	Male	75.67	76.33	91.77	91.63	87.82	95.15	83.76	83.32
	Female	27.47	27.56	10.29	9.87	13.44	5.20	20.54	20.29
	No Gender	0.31	0.47	0.35	0.39	0.27	0.00	0.30	0.35
Country	US	59.62	58.66	29.70	26.28	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	1.79	1.18	10.06	10.34	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
Professional Field	Film/TV	39.19	37.93	3.16	1.89	10.72	1.32	20.64	20.75
	Writing	7.40	6.95	36.62	36.39	10.46	6.70	18.41	18.05
	Politics	11.98	10.84	24.02	24.86	36.97	46.61	21.18	20.72
	Science/Tech	3.61	4.71	8.93	7.50	13.67	29.60	5.43	5.54
	Music	24.08	23.44	9.51	10.73	13.56	2.64	16.46	17.38
	Sports	13.69	16.49	1.31	0.32	2.99	1.76	11.03	11.19
	Stage	15.67	14.84	1.49	1.03	1.21	1.23	10.89	10.28
	Artist	1.84	2.47	9.17	9.00	4.58	2.38	5.75	5.93

Table 2: Coverage (% of examples) of demographic values across examples with people in QA datasets. Men dominate, as do Americans.

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) with a 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 χ^2 tests (Plackett, 1983) to see *if* trends exist and using logistic regression to explore those that do.

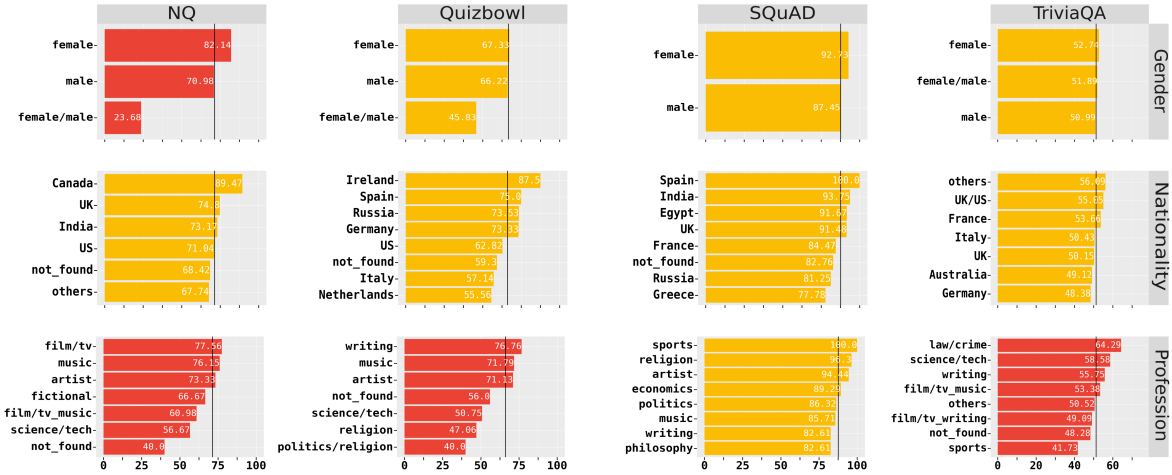

Figure 1: Accuracies split by demographic subsets in our QA datasets' dev fold for all three characteristics compared to average accuracy (vertical line). For each dataset, we only consider examples that have a mention of a person-entity in either the answer, question or document title. Individual plots correspond to a χ^2 test on whether demographic values and accuracy are independent (Section 3.1) with the significant characteristics highlighted in red (p-value < 0.0167).

3.1 Do Demographic Values Affect Accuracy?

The χ^2 test is a non-parametric test of whether two variables are independent. To see if accuracy and characteristics are independent, we apply a χ^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 χ^2 identifies *which* characteristics impact model accuracy, it does not characterize *how*. For instance, χ^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?

Gender	# Correct	# Incorrect	Accuracy
`male`	433	177	70.98%
`female`	161	35	82.14%
`female/male`	9	29	23.68%

Table 3: Illustration of $n \times 2$ contingency table for the χ^2-test: **Gender** in **NQ**, with $n = 3$ values: **`male`**, **`female`** and **`female/male`**. Female entities have a higher accuracy (82%) than male (71%). With two degrees of freedom and $\chi^2 = 53.55$, we get $p = 2.4 \times 10^{-12}$, signaling significance.

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 of characteristics the χ^2 test found significant (thus we exclude all of SQuAD, the nationality characteristic, and gender characteristic for all but NQ). For instance, a question about Abidali Neemuchwala would have features for **`g_male`**, **`o_executive`** but zero for everything else.[5] Real-valued features, **`multi_entities`** and **`multi_answers`**, capture the effect of multiple

[5] Exhaustive list of demographic features in the Appendix.

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_who` checks if the token "who" is in the start of the question. Similarly, `t_what`, `t_when`, and `t_where` capture other entity-types. Questions are also easier if evidence only differs from the question by a couple of words; thus, `q_sim` is the Jaccard similarity between question and evidence tokens. Finally, the binary feature `e_train_count` marks if the person-entities occur in the training data more than twice.

We first drop features with negligible effect on accuracy using LASSO (regularization $\lambda = 1$) by removing zero coefficients. For the 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_sim` and `e_train_count`), those associated with incorrect answers have negative weights, and features without effect will be near zero. Among the `t_wh*` features, `t_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 `multi_answer` (and `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 χ^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

Dataset	Features	Coef	SE	Wald (|W|)	$\mathbb{P}_{Z\sim\mathcal{N}}(\|Z\| > \|W\|)$	
NQ Model Fit: 78.09%	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	⋆⋆
	t_who	+0.773	0.280	2.764	0.0057	⋆⋆⋆
	o_science/tech	-0.715	0.390	1.832	0.0670	⋆
	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	⋆⋆
QB Model Fit: 71.90%	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	0.265	2.522	0.0117	⋆⋆
	o_religion	-0.655	0.362	1.812	0.0700	⋆
	o_writing	+0.402	0.189	2.128	0.0334	⋆⋆
	t_who	+0.363	0.183	1.990	0.0466	⋆⋆
TriviaQA Model Fit: 60.18%	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	⋆

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

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 useful in designing the entity linking and attribute extraction pipeline.

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 *Proceedings of International Conference on Computational Linguistics*, Beijing, China.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion 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 Association for Computational Linguistics*, Online.

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, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of Empirical Methods in Natural Language Processing*, Online.

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? In *Proceedings of the Association for Computational Linguistics*.

Candace Ross, Boris Katz, and Andrei Barbu. 2021. Measuring social biases in grounded vision and language embeddings. In *Conference of the North American Chapter of the Association for Computational Linguistics*, Online.

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? In *Proceedings of Empirical Methods in Natural Language Processing*.

Chenglei Si, Chen Zhao, and Jordan Boyd-Graber. 2021. What's in a name? answer equivalence for open-domain question answering. In *Emperical 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

secretary, fictional soldier, fictional space traveller, fictional taxi driver, fictional vigilante, fictional waitperson, fictional writer

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

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, philosopher of science

Polymath: polymath

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.[6]

- `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 single 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 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.

[6] QB questions often start with "For 10 points, name this writer *who*..."

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: -4135209844918483842
Q: *who carried the us flag in the 2014 olympics*
A: Todd Lodwick
A: Julie Chu
id: 8838716539218945006
Q: *who says that which we call a rose*
A: William Shakespeare
A: Juliet
id: -6197052503812142206
Q: *who has won the most superbowls as a player*
A: Charles Haley
A: Tom Brady
id: -2840415450119119129
Q: *who started the guinness book of world records*
A: Hugh Beaver
A: Norris and Ross McWhirter
id: 6997422338613101186
Q: *who played the nurse on andy griffith show*
A: Langdon
A: Julie Adams
id: -7064677612340044331
Q: *who wrote the song if i were a boy*
A: BC Jean
A: Toby Gad
id: 3248410603422198181
Q: *who conducted the opening concert at carnegie hall*

Dataset	Features	Coef	SE	Wald (|Wl)	$\mathbb{P}_{Z\sim\mathcal{N}}(\|Z\| > \|W\|)$	
NQ **Model Fit:** **78.09%**	**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**	⋆⋆
	t_who	**+0.773**	**0.280**	**2.764**	**0.0057**	⋆⋆⋆
	o_law/crime	+0.729	0.738	0.989	0.3227	
	o_science/tech	**-0.715**	**0.390**	**1.832**	**0.0670**	⋆
	t_where	-0.695	0.645	1.078	0.2811	
	multi_entities	**-0.678**	**0.342**	**1.979**	**0.0478**	⋆⋆
	t_what	+0.505	0.415	1.216	0.2240	
	o_stage	-0.444	0.607	0.732	0.4640	
	g_female	+0.427	0.489	0.873	0.3829	
	o_executive	+0.425	0.684	0.622	0.5341	
	q_sim	**+0.406**	**0.210**	**1.934**	**0.0531**	⋆
	e_train_count	**+0.353**	**0.178**	**1.979**	**0.0479**	⋆⋆
	o_others	+0.346	0.443	0.781	0.4346	
	o_music	+0.310	0.259	1.200	0.2302	
	o_fictional	-0.256	0.445	0.576	0.5644	
	o_politics	-0.197	0.304	0.646	0.5185	
	o_film/tv	+0.142	0.226	0.629	0.5294	
	t_when	-0.134	0.378	0.354	0.7233	
	g_male	-0.126	0.508	0.249	0.8033	
QB **Model Fit:** **71.90%**	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**	⋆⋆⋆⋆
	t_what	-0.734	0.457	1.605	0.1085	
	o_translation	+0.671	1.439	0.466	0.6412	
	o_science/tech	**-0.667**	**0.265**	**2.522**	**0.0117**	⋆⋆
	o_religion	**-0.655**	**0.362**	**1.812**	**0.0700**	⋆
	t_where	+0.545	0.615	0.885	0.3760	
	o_history	-0.484	0.436	1.109	0.2673	
	o_writing	**+0.402**	**0.189**	**2.128**	**0.0334**	⋆⋆
	o_not_found	-0.382	0.349	1.093	0.2744	
	o_fictional	-0.368	0.540	0.681	0.4957	
	t_who	**+0.363**	**0.183**	**1.990**	**0.0466**	⋆⋆
	o_artist	+0.194	0.268	0.721	0.4709	
	o_music	+0.123	0.249	0.492	0.6225	
	o_politics	-0.113	0.202	0.560	0.5756	
	o_philosophy	-0.082	0.259	0.315	0.7530	
TriviaQA **Model Fit:** **60.18%**	**bias**	**-1.066**	**0.114**	**9.353**	**0.0000**	⋆⋆⋆⋆
	o_education	-0.665	0.534	1.245	0.2132	
	o_economics	+0.503	0.579	0.869	0.3848	
	o_philosophy	-0.469	0.342	1.370	0.1706	
	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_science/tech	+0.219	0.151	1.452	0.1466	
	o_stage	-0.212	0.257	0.824	0.4098	
	o_politics	**-0.208**	**0.095**	**2.177**	**0.0295**	⋆⋆
	o_writing	**+0.192**	**0.098**	**1.955**	**0.0506**	⋆
	o_sports	-0.183	0.117	1.570	0.1164	
	o_film/tv	+0.150	0.092	1.626	0.1039	
	t_when	-0.146	0.299	0.488	0.6258	
	o_not_found	-0.143	0.205	0.699	0.4845	
	o_executive	-0.130	0.207	0.625	0.5323	
	o_others	-0.101	0.152	0.665	0.5058	
	multi_entities	-0.099	0.077	1.290	0.1971	
	e_train_count	+0.081	0.075	1.079	0.2804	
	o_music	+0.022	0.097	0.223	0.8237	
	t_what	+0.018	0.134	0.132	0.8947	

Table 5: Influential features revealed through Logistic Regression Analysis (Sec 3.2) over the demographic characteristics deemed significant through the χ^2 test (Figure 1). We report the highly influential features with significance of p-value < 0.1, both positive (blue) and negative (red), and **bold** the highly significant ones (p-value < 0.05). Number of ⋆ in the last column represents the significance level of that feature.

A: Walter Damrosch
A: Pyotr Ilyich Tchaikovsky
id: -3772952199709196386
Q: *who founded amazon where is the headquarters of amazon*
A: founded by Jeff Bezos
A: based in Seattle , Washington
id: 4053461415821443645
Q: *who wrote song what a friend we have in jesus*
A: Joseph M. Scriven
A: Charles Crozat Converse
id: -5670674709553776773
Q: *who sings the theme song for the proud family*
A: Solange Knowles
A: Solange Knowles
id: 2978779480736570480
Q: *days of our lives cast doug and julie*
A: Bill Hayes
A: Susan Seaforth
id: 6173192803639008655
Q: *who has appeared in the most royal rumbles*
A: Isaac Yankem / " Diesel " / Kane
A: Shawn Michaels
id: 7561389892504775773
Q: *who wrote the song stop the world and let me off*
A: Carl Belew
A: W.S. Stevenson
id: -8366545547296627039
Q: *who wrote the song photograph by ringo starr*
A: Ringo Starr

A: George Harrison
id: -5674327280636928690
Q: *who sings you're welcome in moana credits*
A: Lin - Manuel Miranda
A: Jordan Fisher
id: -2432292250757146771
Q: *who wrote the song i hate you i love you*
A: Garrett Nash
A: Olivia O'Brien
id: -3632974700795137148
Q: *who is the owner of reading football club*
A: Xiu Li Dai
A: Yongge Dai
id: 7163132803738849961
Q: *who played guitar on my guitar gently weeps*
A: Eric Clapton
A: George Harrison
id: 1318031841813121387
Q: *who sang the theme song to that 70s show*
A: Todd Griffin
A: Cheap Trick
id: 1393634180793653648
Q: *who came up with the initial concept of protons and neutrons*
A: Werner Heisenberg
A: Dmitri Ivanenko
id: 9134704289334516617
Q: *who missed the plane the day the music died*
A: Waylon Jennings
A: Tommy Allsup

id: 8466196474705624263
Q: *who was running as vice president in 1984*
A: Congresswoman Ferraro
A: Vice President George H.W. Bush
id: 5579013873387598720
Q: *who has won the canada open women's doubles*
A: Mayu Matsumoto
A: Wakana Nagahara
id: 5584540254904933863
Q: *who sang what are we doing in love*
A: Dottie West
A: Kenny Rogers
id: -8677459248394445003
Q: *who is hosting e live from the red carpet*
A: Ryan Seacrest
A: Giuliana Rancic
id: -1342189058950802702
Q: *who made the poppies at tower of london*
A: Paul Cummins
A: Tom Piper
id: 6014950976264156000
Q: *who sang never gonna let you go*
A: Joe Pizzulo
A: Leeza Miller
id: -8052136860650205450
Q: *who wrote the song rainy days and mondays*
A: Roger Nichols
A: Paul Williams
id: 7903911150166287814
Q: *what position did doug peterson play in the nfl*
A: quarterback
A: holder on placekicks
id: 583026970021621830
Q: *who invented the first home video security system*
A: Marie Van Brittan Brown
A: her husband Albert Brown
id: 5427679691711111925
Q: *who were the two mathematicians that invented calculus*
A: Isaac Newton
A: Gottfried Leibniz
id: -9163844183450408581
Q: *nba record for most double doubles in a season*
A: Tim Duncan leads the National Basketball Association (NBA) in the points - rebounds combination with 840
A: John Stockton leads the points - assists combination with 714
id: -8109367537690343895
Q: *who were the twins that played for kentucky*
A: Andrew Michael Harrison
A: Aaron Harrison
id: 4784420206031467202
Q: *who wrote he ain't heavy he's my brother lyrics*
A: Bobby Scott
A: Bob Russell
id: 4136958282795887427
Q: *who opens the church of the holy sepulchre*
A: the Nusaybah family
A: the Joudeh Al - Goudia family
id: -2610209560699528896
Q: *who is the writer of 50 shades of grey*
A: Erika Mitchell Leonard
A: E.L. James
id: 8968036245733884389
Q: *when did stephen curry won the mvp award*
A: 2015 ,
A: 2015
id: -1899514742808499173
Q: *who are nominated for president of india 2017*
A: Ram Nath Kovind
A: Meira Kumar
id: -3019484115332998709
Q: *what movie is count on me by bruno mars in*
A: A Turtle 's Tale : Sammy 's Adventures
A: Diary of a Wimpy Kid : The Long Haul
id: 810060125994185205
Q: *who was the first to say i'm going to disney world*
A: Dick Rutan
A: Jeana Yeager
id: 339027965927992295
Q: *who sings the whiskey ain't workin anymore*
A: Travis Tritt
A: Marty Stuart
id: 5995814638252489040
Q: *who played scotty baldwins father on general hospital*
A: Peter Hansen
A: Ross Elliott
id: 3723628014502752965
Q: *who wrote cant get you out of my head lyrics*
A: Cathy Dennis
A: Rob Davis
id: 3886074985605209321
Q: *who sings find out who your friends are with tracy lawrence*
A: Tim McGraw
A: Kenny Chesney
id: 3624266518328727040
Q: *who invented the printing press and what year*
A: Johannes Gutenberg
A: circa 1439
id: -4951004239400083779
Q: *who plays chris grandy in 13 going on 30*
A: Jim Gaffigan
A: Alex Black
id: 2672721743911117185
Q: *who developed a set of postulates to prove that specific microorganisms cause disease*
A: Robert Koch
A: Friedrich Loeffler
id: 2166092801797515500
Q: *who is the director of taarak mehta ka ooltah chashmah*
A: Harshad Joshi
A: Malav Suresh Rajda
id: -3389723371168293793
Q: *who has the most olympic medals in figure skating*
A: Tessa Virtue
A: Scott Moir
id: -8391680223788694572
Q: *who wrote if i were a boy reba or beyonce*
A: BC Jean
A: Toby Gad
id: 1070572237499172286
Q: *who wrote the song after you've gone*
A: Turner Layton
A: Henry Creamer
id: 2343902375984110832
Q: *who does the voice of mickey mouse on mickey mouse clubhouse*
A: Wayne Allwine
A: Bret Iwan
id: 7013863939803495694
Q: *who sings love me tender in princess diaries 2*
A: Norah Jones
A: Adam Levy
id: 4925057086725798331
Q: *who wrote yakkity yak don't talk back*
A: Jerry Leiber
A: Mike Stoller
id: 647605647914971565
Q: *who wrote lyrics for phantom of the opera*
A: Charles Hart
A: Richard Stilgoe
id: -6371603500131574271
Q: *who sings somebody's watching me with michael jackson*
A: Rockwell
A: Jermaine Jackson
id: -4036503601399675973
Q: *when did michael jordan return to the nba*
A: 1995
A: 2001
id: 4323871331649279373
Q: *who invented the printing press and in what year*
A: Johannes Gutenberg
A: 1440
id: 7234277123646852447
Q: *who sings war don't let me down*
A: American production duo The Chainsmokers
A: vocals of American singer Daya
id: 4245798066923223457
Q: *who has the most all star mvp awards*
A: Bob Pettit
A: Kobe Bryant
id: -3585157729928173881
Q: *who plays hulk in the thor and avengers series of movies*
A: Fred Tatasciore
A: Rick D. Wasserman
id: -7892904540301629325
Q: *who wrote the song going to kansas city*
A: Jerry Leiber
A: Mike Stoller
id: 1838851770314085590
Q: *who plays sheila carter on the bold and the beautiful*
A: Kimberlin Brown
A: Michelle Stafford

B.2.2 TriviaQA multi-answer examples:

We randomly sample 100 examples from TriviaQA where questions had multiple answers.

id: sfq_6110
Q: *On which island in the North Sea did both St Aidan and St Cuthbert live?*
A: Lindisfarne
A: LINDISFARNE
id: tc_1008
Q: *To the nearest two, how many tennis Grand Slam titles did Jimmy Connors win?*
A: 10
A: ten
id: sfq_26211
Q: *In the TV series Doctor Who, who was the creator of the Daleks and arch enemy of the Doctor?*
A: Davros
A: Creator of the Daleks
id: sfq_22212
Q: *In which book of the bible is the story of Samson and Delilah?*
A: Judge (disambiguation)
A: Judges
id: bt_1538
Q: *What is cartoon character Mr. Magoo's first name*
A: Quincy (disambiguation)
A: Quincy

id: qz_2444
Q: *What is Robin Williams character called in Good Morning Vietnam?*
A: Adrian
id: sfq_22693
Q: *What was the first name of the jazz trombonist Kid Ory?*
A: Eadweard
A: Edward
id: qw_1606
Q: *Which of Queen Elizabeth's children is the lowest in succession to (i.e. furthest away from) the throne?*
A: Anne
A: Ann (name)
id: odql_5503
Q: *"Which radio comedian's catchphrase was ""daft as a brush""?"*
A: KEN PLATT
A: Ken Platt
id: qg_2992
Q: *According to Sammy Haggar, what can't he drive?*
A: 55
A: fifty-five
id: qf_3440
Q: *What was Grace Darling's father's job?*
A: Lighthouse-keeper
A: Lighthouse keeper
id: qw_12369
Q: *"What year did Jean-Francois Champollion publish 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: one thousand, eight hundred and twenty-two
A: 1822
id: qw_4143
Q: *What is the title of the most famous painting by Franz Hals?*
A: Laughing Cavalier
A: The Laughing Cavalier
id: qb_2647
Q: *What is the title of the 1944 film starring Barbara Stanwyck as the wife who seduces an insurance salesman into killing her husband?*
A: Double indemnity (disambiguation)
A: Double Indemnity
id: sfq_22920
Q: *Who was the choreographer of the dance troupe Hot Gossip?*
A: Arlene Philips
A: Arlene Phillips
id: tc_719
Q: *River Phoenix died during the making of which movie?*
A: Dark Blood (film)
A: Dark Blood
id: sfq_19457
Q: *Who won the first ever boxing gold for women? She shares her surname with two US Presidents.*
A: Nicola Adams
A: Adams, Nicola
id: sfq_8996
Q: *Actor Norman Painting died in November 2009, which part in a log running radio series did he make his own?*
A: PHIL ARCHER
A: Phil Archer
id: dpql_6111
Q: *Which Jersey-born actor played Superman in Man of Steel?*
A: Henry Cavill
A: Henry William Dalgliesh Cavill
id: qz_2135
Q: *Name the game show, presented by Leslie Grantham and Melinda Messenger, where contestants were set physical and mental challenges?*
A: Fort Boyard (disambiguation)
A: Fort Boyard
id: odql_3205
Q: *Who wrote the novel 'The Beach' on which the film was based?*
A: Alex Garland
A: ALEX GARLAND
id: qz_2999
Q: *In what year did Edward Vlll abdicate?*
A: one thousand, nine hundred and thirty-six
A: 1936
id: tc_723
Q: *Which artist David was born in Bradford UK?*
A: Hockney
A: David Hockney
id: odql_3708
Q: *Three Liverpool players were in the 1966 England World Cup winning squad. Roger Hunt and Ian Callaghan were two – who was the third?*
A: Gerry Byrne
A: Gerry Byrne (disambiguation)
id: qw_11151
Q: *Which artist has a daughter and two sons with Jane Asher, whom he married in 1981?*
A: Gerald Anthony Scarfe
A: Gerald Scarfe
id: qb_3652
Q: *Who wrote the novel 'The Eagle Has landed'?*
A: Harry Patterson
A: Jack Higgins
id: odql_14683
Q: *Who presents the BBC quiz show 'Perfection'?*
A: Nick Knowles
A: NICK KNOWLES
id: sfq_9464
Q: *Who succeeded Brian Epstein as manager of The Beatles?*
A: Allan Klein
A: Allen Klein
id: wh_2615
Q: *In which year did both T-Rex's Marc Bolan and Elvis Presley die ?*
A: 1977
A: one thousand, nine hundred and seventy-seven
id: sfq_9018
Q: *Who played Hotlips Houlihan in the 1972 film MASH?*
A: Sally Kellerman
A: SALLY KELLERMAN
id: qz_1516
Q: *Who bought Chelsea football club for £1 in 1982?*
A: Ken Bates
A: Kenneth Bates
id: sfq_23289
Q: *What was the middle name of the author William Thackeray?*
A: Makepeace
A: MAKEPEACE
id: sfq_7589
Q: *What was the name of the older brother of Henry 8th?*
A: Arthur
A: Arthur (name)
id: odql_10316
Q: *Which actor played 'Hadley', in the TV series of the same name?*
A: GERALD HARPER
A: Gerald Harper
id: sfq_12933
Q: *Operation Barbarossa, Hitler invades Russia.*
A: one thousand, nine hundred and forty-one
A: 1941
id: qw_5050
Q: *"Which Italian nobel prize winner (1934) wrote novels such as ""Mal Gioconda"" and switched to writing plays in 1910?"*
A: Pirandello
A: Luigi Pirandello
id: bt_2403
Q: *What was the name of the driver of the mail train robbed by the great train robbers*
A: Jack Mills (train driver)
A: Jack Mills
id: sfq_2189
Q: *What was the name of the private eye played by Trevor Eve on TV in the '70s?*
A: Shoestring (TV series)
A: Eddie Shoestring
id: qb_5431
Q: *Brazilian football legend Pele wore which number on his shirt?*
A: 10
A: ten
id: qb_4726
Q: *Michael J Fox travels back to which year in the Wild West in the 1990 film 'Back To The Future Part III'?*
A: one thousand, eight hundred and eighty-five
A: 1885
id: odql_8275
Q: *Later a 'Blue Peter' presenter, who played 'Steven Taylor', an assistant to William Hartnell's 'Doctor Who'?*
A: PETER PURVES
A: Peter Purves
id: sfq_962
Q: *Which city was the subject of the 1949 song 'Dirty Old Town' by Ewan McColl?*
A: Salford
A: Salford (disambiguation)
id: tc_1348
Q: *In the late 60s Owen Finlay MacLaren pioneered what useful item for parents of small chldren?*
A: Baby Buggy
A: Baby buggy
id: qw_12732
Q: *General Franco, the Spanish military general, was head of state of Spain from October 1936 following the Spanish Civil War, until when?*
A: 1975
A: one thousand, nine hundred and seventy-five
id: bt_4495
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
A: Ronald %22Buster%22 Edwards
id: sfq_20394
Q: *Which TV presenter, who died in February 2013, was for over 20 years the host of 'Mr & Mrs'?*
A: Derek Batey
A: Derek Beatty
id: odql_12918
Q: *Which British political party leader is MP for Westmorland and Lonsdale?*
A: Tim Farron
A: Timothy Farron
id: odql_13785
Q: *Who wrote the lyrics for 'Sing', written to celebrate the Queen's Diamond Jubilee?*
A: Gary Barlow
A: GARY BARLOW
id: sfq_20830
Q: *Which top National Hunt trainer's establishment is based at Seven Barrows?*
A: NICKY HENDERSON
A: Nicky Henderson
id: wh_2133
Q: *Which T.V. Quiz show host used the catchphrase :- If its' up there, I'll give you the money myself ?*
A: LES DENNIS
A: Les Dennis

id: sfq_15663
Q: *The 27 episodes of which sitcom featuring Julia Mckenzie, Anton Rodgers and Ballard Berkley were first broadcast in the 1980s?*
A: Fresh Fields (TV series)
A: Fresh Fields
id: odql_4871
Q: *When US President James Garfield was shot in Washington DC in July 1881, what was he doing?*
A: WAITING FOR A TRAIN
A: Waiting for a Train
id: bb_6592
Q: *Which artist was born in Bradford in 1937?*
A: Hockney
A: David Hockney
id: qw_10270
Q: *Argentina invaded UK's Falkland Islands, Israel invaded Southern Lebanon, Canada became officially independent of the UK, Leonid Brezhnev, leader of the USSR, died, all in what year?*
A: one thousand, nine hundred and eighty-two
A: 1982
id: bt_4206
Q: *Who was the first woman to be seen on Channel 4*
A: Carol Vorderman
A: Carol Voderman
id: qw_8871
Q: *Lieutenant General James Thomas Brudenell, who commanded the Light Brigade of the British Army during the Crimean War, was the 7th Earl of what?*
A: Cardigan
A: Cardigan (disambiguation)
id: sfq_13639
Q: *Which model village did Samuel Greg build to house workers at his nearby Quarry Bank Mill?*
A: Styal
A: STYAL
id: tc_812
Q: *Who was the defending champion when Martina Navratilova first won Wimbledon singles?*
A: Virginia Wade
A: Sarah Virginia Wade
id: sfq_11790
Q: *Opened in 1963, which London nightclub did Mark Birley name after his then wife?*
A: Annabel's
A: ANNABELS
id: qw_8397
Q: *In 1995, Steffi Graf became the only tennis player to have won each of the four grand slam events how many times?*
A: four
A: 4
id: dpql_3151
Q: *On which river does Ipswich stand?*
A: Orwell (disambiguation)
A: Orwell
id: qw_14634
Q: *"Which Bob Dylan song begins ""You got a lotta nerveTo say you are my friend. When I was down, You just stood there grinning""?"*
A: Positively Fourth Street
A: Positively 4th Street
id: dpql_1801
Q: *Nick Begs was lead singer with which 80's pop band?*
A: Kaja Googoo
A: Kajagoogoo
id: qw_16011
Q: *In 1483, who was appointed the first grand inquisitor of the Spanish Inquisition?*
A: Torquemada (disambiguation)
A: Torquemada
id: qw_1933
Q: *What remake of a British science-fiction serial broadcast by BBC Television in the summer of 1953 was staged live by BBC Four in 2005 with actors Jason Flemyng, Mark Gatiss, Andrew Tiernan, Indira Varma, David Tennant and Adrian Bower?*
A: Quatermass experiment
A: The Quatermass Experiment
id: odql_2323
Q: *Which 2009 film is a biopic of John Lennon?*
A: 'NOWHERE BOY'
A: Nowhere Boy
id: bb_522
Q: *'The Battle of Trafalgar' is the work of which British painter?*
A: Joseph Turner
A: Joseph Turner (disambiguation)
id: qw_4463
Q: *Who discovered the two moons of Mars in 1877?*
A: Asaph Hall
A: Asaph Hall III
id: qz_1111
Q: *Which brand of beer does Homer Simpson drink regularly?*
A: Duff
A: Duff (disambiguation)
id: wh_8
Q: *In the novel 'Treasure Island' name the pirate shot dead by Jim Hawkins in the rigging of the Hispaniola*
A: Israel Hands
A: ISRAEL HANDS
id: qg_3884
Q: *Dow Constantine and Susan Hutchinson are currently running for was position?*
A: King County executive
A: King County Executive
id: sfq_25940
Q: *Which actor played the part of Ross Poldark in the BBC's mid 1970's television series?*
A: Robin Ellis
A: ROBIN ELLIS
id: odql_12777
Q: *For which 1960 film did Billy Wilder become the first person to win three Oscars for the same film?*
A: The Apartment
A: The apartment
id: bb_4540
Q: *Famous for 'Die Welt als Wille und Vorstellung', Arthur Schopenhauer (1788-1860) was a German?*
A: Philosophers
A: Philosopher
id: qb_8589
Q: *What is the nickname of the frontiersman Nathaniel Poe, played by Daniel Day Lewis, in the 1992, film 'The Last of the Mohicans'?*
A: Hawkeye
A: Hawkeye (disambiguation)
id: bt_2852
Q: *Who played the part of Tina Seabrook in Casualty*
A: Claire Woodrow
A: Claire Goose
id: wh_557
Q: *Who duetted with Syd Owen on the single Better Believe It, which was released as part of the Children in Need appeal in 1995 ?*
A: PATSY PALMER
A: Patsy Palmer
id: odql_3979
Q: *What is the name of the character played by Nicole Kidman in the film 'Moulin Rouge'?*
A: Satine
A: 'SATINE'
id: odql_10365
Q: *Which British girl won the Women's Junior Singles title at Wimbledon this year (2008)?*
A: LAURA ROBSON
A: Laura Robson
id: qf_1735
Q: *In what year did Elvis Presley and his parents move from Tupelo to Memphis?*
A: one thousand, nine hundred and forty-eight
A: 1948
id: tc_1468
Q: *What was Pete Sampras seeded when he won his first US Open?*
A: twelve
A: 12
id: qf_2679
Q: *Who on TV has played a scarecrow and a Time Lord?*
A: John Pertwee
A: Jon Pertwee
id: sfq_11844
Q: *In which year was Olaf Palme assassinated and the Chernobyl nuclear power station exploded?*
A: 1986
A: one thousand, nine hundred and eighty-six
id: qf_3578
Q: *Cassandra was the pseudonym of which writer in the Daily Mirror?*
A: William Neil Connor
A: William Connor
id: qz_3898
Q: *How many times did Steffi Graf win the Ladies Singles at Wimbledon?*
A: seven
A: 7
id: qw_6782
Q: *What is the disease that Stephen Hawking has?*
A: Motor neuron disease
A: Motor neuron diseases
id: qw_15255
Q: *"How many films were made by director Sir Peter Jackson from Tolkien's short book, ""The Hobbit""?"*
A: 3
A: three
id: sfq_7168
Q: *Who invented the wind-up radio?*
A: Trevor Bayliss
A: TREVOR BAYLISS
id: sfq_14433
Q: *In Pride and Prejudice what was the first name of Mr Darcy?*
A: Fitzwilliam (disambiguation)
A: Fitzwilliam
id: odql_8230
Q: *Which single by 'Leapy Lee' reached number two in the UK charts in 1968?*
A: 'LITTLE ARROWS'
A: Little Arrows
id: dpql_1416
Q: *Whose is the first tale in Chaucer's Canterbury Tales?*
A: The Knight
A: Knight (disambiguation)
id: qw_1672
Q: *Which womens squash player won the World Open four times (1985, 1987, 1990 & 1992) and the British Open eight times?*
A: Susan Devoy
A: Susan Elizabeth Anne Devoy
id: qz_832
Q: *Who wrote the novels About A Boy, How To Be Good and High Fidelity?*
A: Nick Hornby
A: Hornby, Nick
id: sfq_22884
Q: *Which TV series was about a pop group called 'Little Ladies' featuring Charlotte Cornwell, Julie Covington and Rula Lenska?*
A: Rock Follies
A: Rock Follies of '77

id: odql_10746
Q: *Who wrote the 1951 novel 'The Caine Mutiny'?*
A: HERMAN WOUK
A: Herman Wouk
id: bb_285
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
A: four hundred and fifty-one
id: odql_7290
Q: *Who was the driver of the limousine at the time of Diana Princess of Wales' death?*
A: HENRI PAUL
A: Henri Paul
id: sfq_4368
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
A: Mustique
id: odql_5476
Q: *Which pop star had the real name of Ernest Evans?*
A: Chubby Checker
A: 'CHUBBY CHECKER'
id: tc_980
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