

# What Do NLP Researchers Believe? Results of the NLP Community Metasurvey

Julian Michael<sup>1</sup>, Ari Holtzman<sup>2</sup>, Alicia Parrish<sup>3\*</sup>, Aaron Mueller<sup>4</sup>, Alex Wang<sup>5\*</sup>,  
Angelica Chen<sup>1</sup>, Divyam Madaan<sup>1</sup>, Nikita Nangia<sup>1</sup>

Richard Yuanzhe Pang<sup>1</sup>, Jason Phang<sup>1</sup>, and Samuel R. Bowman<sup>1,6\*</sup>

<sup>1</sup>New York University, <sup>2</sup>University of Washington, <sup>3</sup>Google,

<sup>4</sup>Johns Hopkins University, <sup>5</sup>Cohere, <sup>6</sup>Anthropic

nlp-metasurvey-admin@nyu.edu

## Abstract

We present the results of the NLP Community Metasurvey. Run from May to June 2022, it elicited opinions on controversial issues, including industry influence in the field, concerns about AGI, and ethics. Our results put concrete numbers to several controversies: For example, respondents are split in half on the importance of artificial general intelligence, whether language models understand language, and the necessity of linguistic structure and inductive bias for solving NLP problems. In addition, the survey posed *meta*-questions, asking respondents to predict the distribution of survey responses. This allows us to uncover *false sociological beliefs* where the community’s predictions don’t match reality. Among other results, we find that the community greatly overestimates its own belief in the usefulness of benchmarks and the potential for scaling to solve real-world problems, while underestimating its belief in the importance of linguistic structure, inductive bias, and interdisciplinary science.

## 1 Introduction

What do NLP researchers think about NLP? Are we devoting too many resources to scaling up? Do language models understand language? Is the traditional paradigm of model benchmarking still tenable? What models are ethical for researchers to build and release?

These questions and many more are actively debated in the research community, and views on them are a major factor in deciding what work gets done. Understanding the prevalence of different views on these issues is valuable for understanding the trajectory of NLP research and the structure of the field. In addition, communication among researchers often rests on *sociological beliefs* about these questions: what people think people think. Getting these sociological beliefs wrong can slow

down communication and lead to wasted effort, missed opportunities, and needless fights. For example, researchers might spend time and effort defending a widely-believed position which they mistakenly think is controversial, or might fail to argue effectively if they appeal to premises which they believe are well-established but are actually contentious or unpopular.

Many of the ways that the NLP research community gets to know itself—invited talks, panel discussions, social media, etc.—are biased, for example through self-selection towards similar people and amplification of already-prominent and controversial voices. This makes it difficult to get a sense of the community’s beliefs as a whole. For these reasons, we believe it is worth trying to objectively assess NLP researchers’ views on controversial issues. So from May to June 2022, we conducted the **NLP Community Metasurvey**. We present stances, such as *Currently, the field focuses too much on scaling up machine learning models (Q5-1)*, and ask respondents whether they agree or disagree. Then we ask them to *predict what percent of respondents who will agree*. This gives us insight into the community’s object-level beliefs as well as its sociological beliefs, and allows us to identify where the two may be misaligned. This work is directly inspired by the PhilPapers Surveys (Bourget and Chalmers, 2014, 2020), an initiative by philosophers to assess the philosophy community’s beliefs about current topics in their field and their sociological beliefs about their professional community.

The rest of this document reports the methodology and results of the NLP Community Metasurvey. Our results are *descriptive*, not *prescriptive*, as these issues cannot be resolved by majority vote. By necessity, we are covering a subjectively chosen set of questions and reducing many complex issues into simplified scales, but we hope that these results can create common ground for fruitful discussion among NLP researchers.

\*Work done while at NYU.

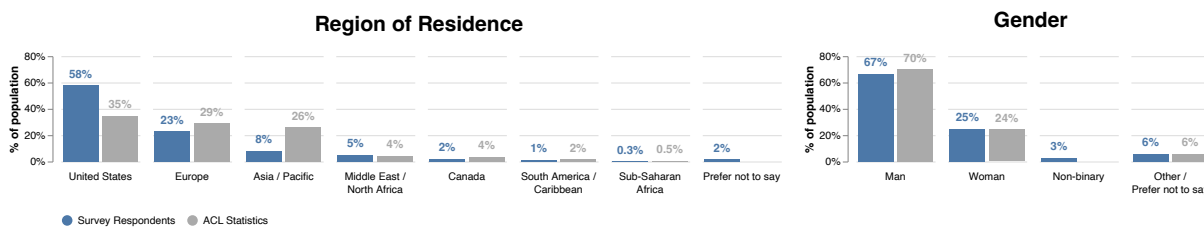


Figure 1: Basic demographics of survey respondents, compared to available statistics from the ACL. For region, we compare to ACL memberships as of summer 2021. For gender, we use the publicly available statistics from attendance at ACL 2017 in Vancouver (which lacked a specific “non-binary” category).

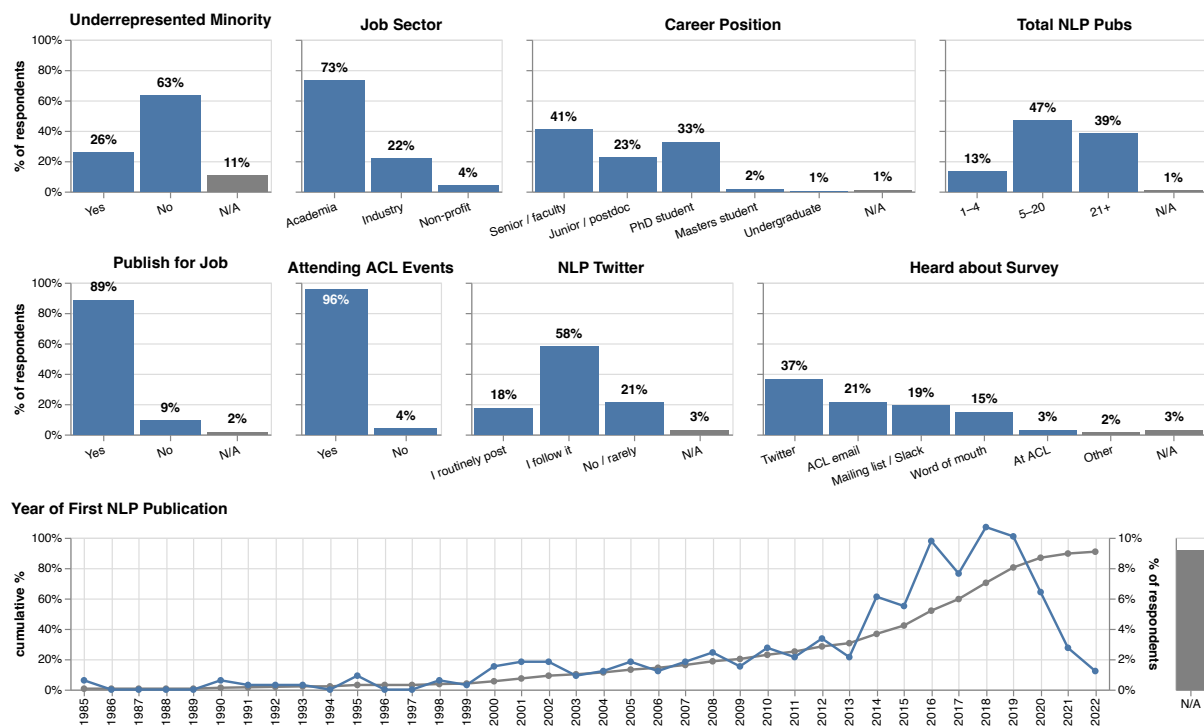


Figure 2: Other demographic information. Skipped questions or “Prefer not to say” answers are listed as “N/A.” Full question text and results are in [Appendix C](#).

## 2 Methodology

**Survey Construction** The survey questions are shown in Figures 3 and 4. We present them in thematic sections (e.g., *State of the Field*), phrased as statements expressing a certain opinion (e.g., *Q1-1. Private firms have too much influence in guiding the trajectory of the field*). Respondents answer on a 4-point scale of **AGREE**, **WEAKLY AGREE**, **WEAKLY DISAGREE**, and **DISAGREE**. We don’t include a neutral middle option because our intent is to push respondents to consider which side they stand on; we instruct them to choose **WEAKLY AGREE** or **WEAKLY DISAGREE** if they have even slight preferences for one side or the other (e.g., “depends, leaning negative”). For cases where they truly cannot make a judgment, we include three

OTHER answers: QUESTION IS ILL-POSED, INSUFFICIENTLY INFORMED ON THE ISSUE, and PREFER NOT TO SAY.

At the end of each section, we ask respondents to predict the percentage of our target population who will either **AGREE** or **WEAKLY AGREE** with each statement. Respondents choose one of five buckets: 0–20%, 20–40%, 40–60%, 60–80%, or 80–100%. These questions can be skipped, but we encourage best guesses even if unsure. Each section has a free-response box for feedback. Survey instructions are reproduced in full in [Appendix E](#).

**Target Demographic: Active Authors in ACL** We define the target population as all co-authors on at least two \*CL papers published in the last three years. This allows us to estimate response bias by

comparing to ACL members (see §3) and provides an objectively defined group for respondents to make predictions about in meta-questions.

**Platform and Distribution** We used NYU Qualtrics to host the survey. Following guidelines set out by the NYU IRB (FY2022-6461), all respondents gave informed consent before beginning the survey and could skip or refuse to answer each question.

We set up a homepage for the survey at <https://nlpsurvey.net> and advertised with Twitter posts, an email via the ACL Member Portal, flyers and stickers at the ACL 2022 conference, personal emails, and other methods (details in Appendix A.3). As an incentive, we committed to donating \$10 for each respondent to one of several non-profits, chosen by the respondent at the end of the survey (see Appendix A.4 for donated amounts).

### 3 Demographics

480 people completed the survey, of which 327 (68%) are in our target demographic, reporting that they co-authored at least 2 ACL publications between 2019 and 2022. Based on the ACL Anthology, 6323 people met this requirement during the survey period, so we have responses from about 5% of the total. For the rest of this paper, we restrict all reported results to this subset. Demographic information is shown in Figures 1 and 2.

**Response Bias** Figure 1 shows location and gender statistics. Survey respondents are mostly men (67%) and mostly from the United States (58%). Comparing to recent official statistics from the ACL<sup>2</sup> suggests that the US is overrepresented (58% > 35%), while Asia/Pacific is underrepresented (8% < 26%). We suspect this is largely due to biases in our survey distribution methods (particularly Twitter and our personal networks). Our gender distribution is roughly comparable to available ACL statistics.<sup>3</sup>

**Career** Figure 2 shows that respondents are mostly from academia (73%, versus 22% in industry), and are relatively senior: 41% report being faculty or senior managers and 39% report >20

<sup>2</sup>[https://www.aclweb.org/adminwiki/images/f/f4/Memberships\\_2021\\_by\\_Country\\_SUMMER.pdf](https://www.aclweb.org/adminwiki/images/f/f4/Memberships_2021_by_Country_SUMMER.pdf)

<sup>3</sup><https://www.aclweb.org/portal/content/acl-diversity-statistics>

publications related to NLP. Information on the subfields of respondents’ work is in Appendix C, Figure 11.

**Twitter Use** Respondents most commonly heard about the survey through Twitter (37%), which is used by a large majority (18% routinely post, 58% follow but don’t often post). While we don’t know what proportion of our target population uses Twitter, it seems likely that this is a large source of response bias. At the same time, the purpose of this survey is partly to study NLP researchers’ perceptions of the NLP community for the purpose of improving the public and scientific discourse. To the extent that these perceptions are formed on Twitter, and the public discourse is carried out on Twitter, our results may be useful even if biased towards Twitter users.

## 4 Results

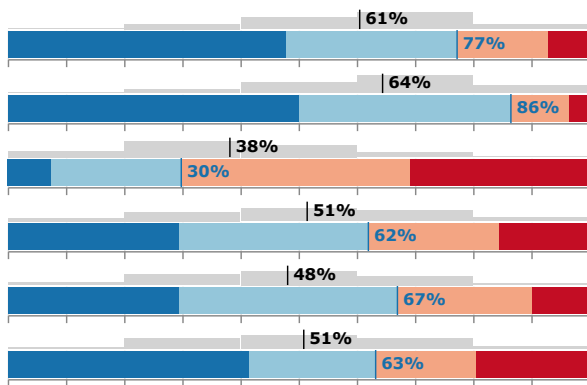
All agree/disagree questions and their responses are shown in Figures 3 and 4. An extended version of this discussion addressing every question is in Appendix D. In the rest of this section, we will discuss some highlights.<sup>4</sup>

**Belief in scaling maximalism is rare and greatly overestimated, but concerns about AGI are not uncommon.** Only a small minority (17%) of respondents agree with the hypothesis that scaling up current systems and methods could solve “practically any important problem” in NLP (Figure 3b, Q2-1), but this view is perceived as much more popular (at 47% predicted agreement). Similarly, a large majority believes that NLP research is focusing too much on scaling up (71% > 58% predicted, Figure 4a, Q5-1). This suggests that the popular discourse around recent developments in scaling up (Chowdhery et al., 2022) may not be reflective of the views of the NLP research community as a whole, which may not have bought into the *Bitter Lesson* (Sutton, 2019)<sup>5</sup> as much as it thinks it has.

Yet, narrow majorities of respondents regard recent progress in large-scale modeling as progress towards artificial general intelligence (AGI; 57%, Figure 3c, Q3-2), and think AGI should be a concern for NLP researchers (58%, Q3-1). So while

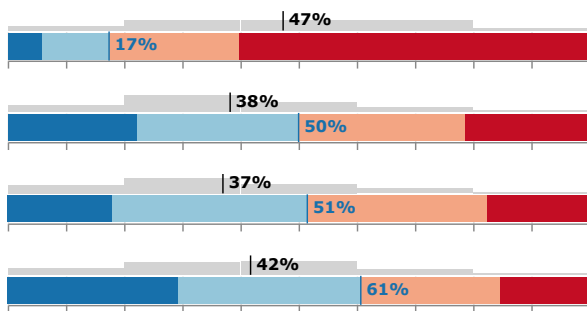
<sup>4</sup>A web interface for exploring the results, including the relationship between questions and demographic variables, is available at <https://nlpsurvey.net/results>.

<sup>5</sup>“General methods that leverage computation are ultimately the most effective, and by a large margin” (Sutton, 2019).



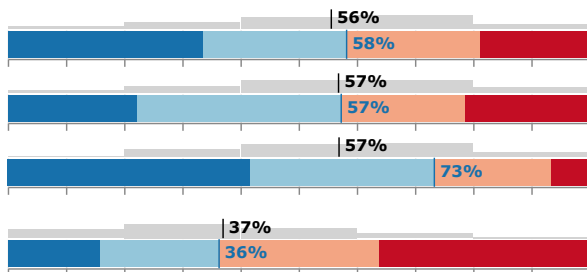
(a) *State of the Field.*

- 1-1. Private firms have too much influence**  
Private firms have too much influence in guiding the trajectory of the field.
- 1-2. Industry will produce the most widely-cited research**  
The most widely-cited papers of the next 10 years are more likely to come out of industry than academia.
- 1-3. NLP winter is coming (10 years)**  
I expect an "NLP winter" to come within the next 10 years, in which funding and job opportunities in NLP R&D fall by at least 50% from their peak.
- 1-4. NLP winter is coming (30 years)**  
I expect an "NLP winter" to come within the next 30 years, in which funding and job opportunities in NLP R&D fall by at least 50% from their peak.
- 1-5. Most of NLP is dubious science**  
A majority of the research being published in NLP is of dubious scientific value.
- 1-6. Author anonymity is worth restricting preprints**  
Author anonymity during review is valuable enough to warrant restrictions on the dissemination of research that is under review.



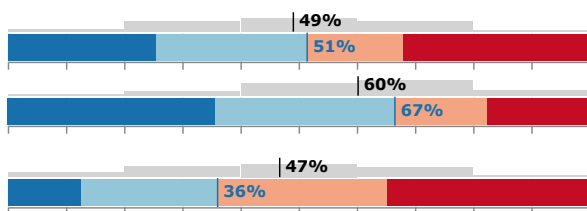
(b) *Scale, Inductive Bias, and Adjacent Fields.*

- 2-1. Scaling solves practically any important problem**  
Given resources (i.e., compute and data) that could come to exist this century, scaled-up implementations of established existing techniques will be sufficient to practically solve any important real-world problem or application in NLP.
- 2-2. Linguistic structure is necessary**  
Discrete general-purpose representations of language structure grounded in linguistic theory (involving, e.g., word sense, syntax, or semantic graphs) will be necessary to practically solve some important real-world problems or applications in NLP.
- 2-3. Expert inductive biases are necessary**  
Expert-designed strong inductive biases (à la universal grammar, symbolic systems, or cognitively-inspired computational primitives) will be necessary to practically solve some important real-world problems or applications in NLP.
- 2-4. Ling/CogSci will contribute to the most-cited models**  
It is likely that at least one of the five most-cited systems in 2030 will take clear inspiration from specific, non-trivial results from the last 50 years of research into linguistics or cognitive science.



(c) *AGI and Major Risks.*

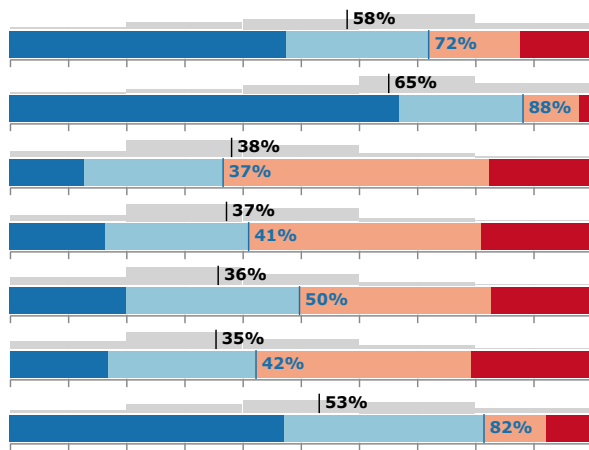
- 3-1. AGI is an important concern**  
Understanding the potential development of artificial general intelligence (AGI) and the benefits/risks associated with it should be a significant priority for NLP researchers.
- 3-2. Recent progress is moving us towards AGI**  
Recent developments in large-scale ML modeling (such as in language modeling and reinforcement learning) are significant steps toward the development of AGI.
- 3-3. AI could soon lead to revolutionary societal change**  
In this century, labor automation caused by advances in AI/ML could plausibly lead to economic restructuring and societal changes on at least the scale of the Industrial Revolution.
- 3-4. AI decisions could cause nuclear-level catastrophe**  
It is plausible that decisions made by AI or machine learning systems could cause a catastrophe this century that is at least as bad as an all-out nuclear war.



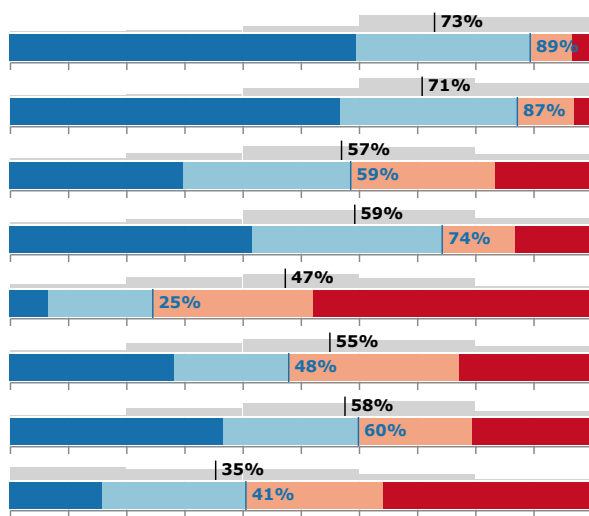
(d) *Language Understanding.*

- 4-1. LMs understand language**  
Some generative model trained only on text, given enough data and computational resources, could understand natural language in some non-trivial sense.
- 4-2. Multimodal models understand language**  
Some multimodal generative model (e.g., one trained with access to images, sensor and actuator data, etc.), given enough data and computational resources, could understand natural language in some non-trivial sense.
- 4-3. Text-only evaluation can measure language understanding**  
We can, in principle, evaluate the degree to which a model understands natural language by tracking its performance on text-only classification or language generation benchmarks.

Figure 3: Responses to questions in sections 1–4 of the survey, colored as **AGREE**, **WEAKLY AGREE**, **WEAKLY DISAGREE**, and **DISAGREE**. On each bar, the lower number (in blue) represents the fraction of respondents who agree with the position out of all those who took a side. The grey bars show the relative proportion of meta-question predictions in each bin (0–20%, 20–40%, etc.), and the upper number (in black) shows the average predicted rate of agreement, computed treating each bin as its midpoint.



(a) Promising Research Programs.



(b) Ethics.

Figure 4: Responses to questions in sections 5–6 of the survey, colored as **AGREE**, **WEAKLY AGREE**, **WEAKLY DISAGREE**, and **DISAGREE**, with the distribution of meta-question answers above each bar in grey.

the most extreme form of scaling maximalism may be unpopular, many researchers seem to think it is worth carefully considering the role of scale in long-term progress.

It is worth considering how views on these issues may have changed since the survey was run in mid-2022: Bubeck et al. (2023) argue that GPT-4 (OpenAI, 2023) is a step towards artificial general intelligence, and scaling training data has brought more performance gains (Anil et al., 2023), but state-of-the-art models also leverage new techniques such as Constitutional AI (Bai et al., 2022).

**Belief in the value of interdisciplinary insights is greatly underestimated, with predicted trend reversals in favor of theoretically-informed approaches to NLP.** A large majority of respon-

5-1. **There's too much focus on scale**  
Currently, the field focuses too much on scaling up machine learning models.

5-2. **There's too much focus on benchmarks**  
Currently, the field focuses too much on optimizing performance on benchmarks.

5-3. **On the wrong track: model architectures**  
The majority of research on model architectures published in the last 5 years is on the wrong track.

5-4. **On the wrong track: language generation**  
The majority of research in open-ended language generation tasks published in the last 5 years is on the wrong track.

5-5. **On the wrong track: explainable models**  
The majority of research in building explainable models published in the last 5 years is on the wrong track.

5-6. **On the wrong track: black-box interpretability**  
The majority of research in interpreting black-box models published in the last 5 years is on the wrong track.

5-7. **We should do more to incorporate interdisciplinary insights**  
Compared to the current state of affairs, NLP researchers should place greater priority on incorporating insights and methods from relevant domain sciences (e.g., sociolinguistics, cognitive science, human-computer interaction).

6-1. **NLP's past net impact is good**  
On net, NLP research has had a positive impact on the world.

6-2. **NLP's future net impact is good**  
On net, NLP research continuing into the future will have a positive impact on the world.

6-3. **It is unethical to build easily-misusable systems**  
It is unethical to build and publicly release a system which can easily be used in harmful ways.

6-4. **Ethical and scientific considerations can conflict**  
In the context of NLP research, ethical considerations can sometimes be at odds with the progress of science.

6-5. **Ethical concerns mostly reduce to data quality and model accuracy**  
The main ethical challenges posed by current ML systems can, in principle, be solved through improvements in data quality/coverage and model accuracy.

6-6. **It is unethical to predict psychological characteristics**  
It is inherently unethical to develop ML systems for predicting people's internal psychological characteristics (e.g., emotions, gender identity, sexual orientation).

6-7. **Carbon footprint is a major concern**  
The carbon footprint of training large models should be a major concern for NLP researchers.

6-8. **NLP should be regulated**  
The development and deployment of NLP systems should be regulated by governments.

dents believe that NLP researchers should place higher priority on incorporating insights from relevant domain sciences such as sociolinguistics, cognitive science, and human-computer interaction (82%, Figure 4a, Q5-7), while greatly underestimating the number of other NLP researchers that share this belief (53% predicted). Relatedly, more respondents than expected believe that theoretically-informed approaches to NLP are necessary for solving real-world NLP problems, either through discrete representations of linguistic structure (50% > 38% predicted, Figure 3b, Q2-2) or expert-designed inductive biases (51% > 31% predicted, Q2-3). A majority also believe that it's likely for one of the five most-cited system in 2030 to take direct inspiration from results from the last

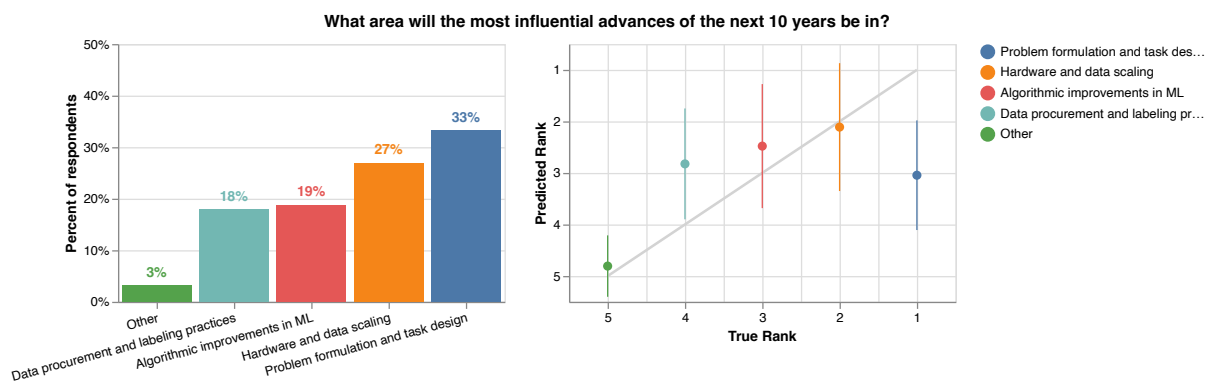


Figure 5: *Likely Sources of Future Advances*. The left shows the distribution of answers, and the right shows predicted ranks (mean and standard deviation) of each answer relative to its true rank.

50 years of linguistics or cognitive science research (61% > 42% predicted, Q2-4). From these results, it seems that many believe there will be a reversal of the current trend of end-to-end modeling with low-bias neural network architectures.

**AGI and LMs understanding language are known controversies.** We find that the community is nearly evenly split on two controversial issues: whether we should be concerned with AGI (58%, Figure 3c, Q3-1), and whether language models understand language (51%, Figure 3d, Q4-1; Bender and Koller, 2020; Michael, 2020; Potts, 2020; Merrill et al., 2021; Bommasani et al., 2021). For both questions, the average meta-response almost perfectly matches the actual percent of people who agree (within 2%), indicating that the community has a good sense that this is a controversial issue.

It is worth acknowledging what this means: these are controversial issues, the community knows that they are controversial, and now (courtesy of this survey) we can know that we know that it’s controversial. Some may believe that AGI is obviously coming soon, and some may believe that it’s obviously ill-defined; some may believe that language models obviously understand language, and some may believe it’s obviously impossible in principle. But for both issues, taking either position for granted in the public discourse or scholarly literature may not be an effective way to communicate to a broad NLP audience. Rather, careful and considered discussion of the issue will be more productive for building common ground.

**The net impact of NLP is believed to be good, but with high stakes.** Large majorities of respon-

dents believe NLP research has had a positive impact on the world (89%, Figure 4b, Q6-1), will have a positive impact in the future (87%, Q6-2), and could plausibly transform society (73%, Figure 3c, Q3-3). Despite this optimism, a substantial minority also foresee plausible risks of a major global catastrophe caused by ML systems (36%, Q3-4). Perhaps surprisingly, 23% of respondents both think the future impact of NLP will be good (Figure 4b, Q6-1) and that AI could plausibly cause a major global catastrophe (Q3-3). Assuming that respondents are consistent about their views, we may understand this to mean that *plausible* catastrophic risk did not necessarily mean *likely* for many respondents, or that the potential upside (e.g., of transforming society) is large enough to outweigh the risk.

#### 4.1 Problem Formulation and Task Design: An Important Frontier?

In addition to the agree/disagree questions constituting most of the survey, we ask respondents where they think the most influential advances of the next 10 years will come from, providing four choices: *Hardware and data scaling*, *Algorithmic improvements in ML*, *Data procurement and labeling practices*, and *Problem formulation and task design*. We also provide an “other” option for people to specify their own answer. As the meta-question, we ask respondents to rank the answers from most popular (1) to least popular (5).<sup>6</sup>

<sup>6</sup>20% of respondents rank “Other” on top (rank 1) for this question, even though very few actually provided it as their answer. These respondents may have ranked the answers backwards by mistake. To correct for this, we reverse the rankings provided by everyone who ranked “Other” first. While not a surefire fix, it doesn’t seem to change any major trends in the results (Appendix A.5, Figure 6).

Q1	Q2	$\rho_s$
Q1-1. Private firms have too much influence.	Q5-1. There's too much focus on scale.	+0.43
Q3-2. Recent progress is moving us towards AGI.	Q4-1. LMs understand language.	+0.42
Q1-5. Most of NLP is dubious science.	Q5-3. On the wrong track: model architectures.	+0.40
Q5-1. There's too much focus on scale.	Q6-7. Carbon footprint is a major concern.	+0.38
Q1-5. Most of NLP is dubious science.	Q4-2. Multimodal models understand language.	+0.38
Q2-2. Linguistic structure is necessary.	Q4-1. LMs understand language.	-0.38
Q4-1. LMs understand language.	Q5-1. There's too much focus on scale.	-0.35
Q1-1. Private firms have too much influence.	Q6-7. Carbon footprint is a major concern.	+0.35
Q2-2. Linguistic structure is necessary.	Q5-7. We should incorporate more interdisciplinary insights.	+0.35
Q2-2. Linguistic structure is necessary.	Q5-1. There's too much focus on scale.	+0.34

Table 1: Top 10 Spearman correlations ( $\rho_s$ ) between pairs of questions from distinct categories (i.e., with distinct first numbers). Correlations  $|\rho_s| > 0.11$  are significant with  $p < 0.05$ .

The results (Figure 5) reveal one surprise: the popularity of the top answer, *Problem formulation and task design*, was greatly underestimated. A plurality of 33% of respondents gave this answer, but it was only ranked first by 12% of people and it ranked third on average in respondents' predictions. Besides this, predictions roughly tracked reality. This suggests that there is a fairly common but underappreciated belief in the NLP community that researchers should be working on new ways of formulating the problems we're trying to solve, and that such work could have high impact.

## 5 Correlation Analysis

This survey provides us the opportunity to examine the relationships between opinions: Which beliefs tend to come together, and which don't? To get a sense of this, we compute pairwise Spearman (rank-order) correlations between pairs of questions (§5.1), and demographic characteristics (§5.2), and perform a clustering analysis on our results using PCA (§5.3).

### 5.1 Question Correlations

We compute Spearman (rank-order) correlations between all of the agree/disagree questions (the full matrix is in Appendix F, Figure 22), ordering the answers [DISAGREE, WEAKLY DISAGREE, OTHER, WEAKLY AGREE, AGREE], where OTHER includes INSUFFICIENTLY INFORMED ON THE ISSUE, QUESTION IS ILL-POSED, and PREFER NOT TO SAY. Unsurprisingly, correlations tend to be stronger between questions in the same section—for example, perspectives on linguistic structure and inductive bias (Figure 3b), AGI (Figure 3c), language understanding (Figure 3d), and NLP's net impact on society (Figure 4b, Q6-1, Q6-2). Be-

yond these, Table 1 shows the highest-magnitude correlations between questions in different sections.

**Concerns about private influence track concerns about scale.** The strongest cross-section correlation is between Q5-1 (there's too much focus on scale) and Q1-1 (private firms have too much influence,  $\rho_s = 0.43$ ). Regarding scale, Q5-1 was also moderately correlated with Q6-7 (carbon footprint is a major concern,  $\rho = 0.38$ ). These correlations suggest that NLP researchers who see the influence of industry as problematic may hold this view in part because of concerns with the large-scale, compute-intensive research paradigm that is spearheaded largely by private firms.

**Believing that LMs understand language is predictive of belief in AGI and the promise of scale.** Agreeing that text-only models can meaningfully “understand” language (Q4-1) is predictive of several other views. Those who agree with Q4-1 are more likely to believe that scaling has moved us towards AGI (Q3-2,  $\rho_s = 0.42$ ) and can solve practically any NLP problem with existing techniques (Q2-1,  $\rho_s = 0.30$ , not shown), and less likely to believe that there is too much focus on scale (Q5-1,  $\rho_s = -0.35$ ), that linguistic structure is necessary to solve important NLP problems (Q2-2,  $\rho_s = -0.38$ ), or that we should do more to incorporate insights from domain sciences (Q5-7,  $\rho_s = -0.30$ , not shown). This suggests the existence of distinct “LM optimist” and “LM pessimist” positions, where people either believe that scaling up could solve most NLP problems and potentially lead to AGI, or they think scaling is overprioritized, AGI is less likely, and we should be focusing more on seeking insights and methods from linguistics,

Demographic	Question	$\rho_s$
Sector: Industry (for-profit)	Q1-1. Private firms have too much influence.	<b>-0.25</b>
Gender: Woman	Q5-7. We should incorporate more interdisciplinary insights.	<b>+0.24</b>
Location: United States	Q4-2. Multimodal models understand language.	<b>+0.22</b>
Location: United States	Q4-1. LMs understand language.	<b>+0.22</b>
Sector: Academia (including students)	Q1-1. Private firms have too much influence.	<b>+0.21</b>
Gender: Man	Q6-2. NLP’s future net impact will be good.	<b>+0.21</b>
Sector: Industry (for-profit)	Q5-7. We should incorporate more interdisciplinary insights.	<b>-0.21</b>
Gender: Woman	Q6-7. Carbon footprint is a major concern.	<b>+0.20</b>
Gender: Woman	Q2-2. Linguistic structure is necessary.	<b>+0.19</b>
Under-represented Minority: Yes	Q3-4. AI decisions could cause nuclear-level catastrophe.	<b>+0.19</b>

Table 2: Top 10 Spearman correlations ( $\rho_s$ ) between membership in a demographic group and answers to questions. Correlations  $|\rho_s| > 0.11$  are statistically significant with  $p < 0.05$ .

Scaling Maximalism (11.7%)		Concern about Fast Progress (5.5%)	
Q4-1. LMs <b>do</b> understand language	+0.36	Q3-1. AGI <b>is</b> important	+0.48
Q4-2. Multimodal models <b>do</b> understand	+0.29	Q3-2. We <b>are</b> stepping to AGI	+0.38
Q6-7. Carbon <b>isn’t</b> a major concern	-0.29	Q6-3. Easy misuse <b>is</b> unethical	+0.28
Q5-1. There <b>isn’t</b> too much focus on scale	-0.28	Q6-7. Carbon <b>is</b> a major concern	+0.26
Q2-2. Linguistic structure <b>isn’t</b> necessary	-0.26	Q3-3. Revolutionary change <b>is</b> plausible	+0.24
Deep Learning Pessimism (5.2%)		Jaded Empiricism (4.0%)	
Q5-5. <b>Wrong</b> track (explainability)	+0.33	Q6-6. <b>Not</b> unethical to predict psych.	-0.37
Q5-6. <b>Wrong</b> track (interpretability)	+0.32	Q1-5. Most NLP <b>is</b> dubious science	+0.30
Q1-5. Most NLP <b>is</b> dubious science	+0.28	Q5-5. <b>Wrong</b> track (explainability)	+0.27
Q3-4. Catastrophic risk <b>is</b> plausible	+0.25	Q5-6. <b>Wrong</b> track (interpretability)	+0.24
Q6-2. NLP <b>isn’t</b> good (future)	-0.23	<b>Has</b> 21+ Pubs.	+0.24

Table 3: The top four components from running PCA on survey responses and demographic data, with informally assigned cluster labels, percent variance explained in parentheses, and the questions/data with the highest magnitude associations per component. We **negate** the statements with negative loadings so the statements for each component correspond to the set of beliefs that vary together.

cognitive science, or other domain sciences. It is worth emphasizing, though, that all of our measured correlations are weak to moderate; people hold diverse sets of opinions and individuals cannot be cleanly split into these camps.

## 5.2 Demographic Correlations

Spearman correlations between demographic variables and agree/disagree questions are shown in Appendix F, Figure 23, with the top correlations by magnitude in Table 2. We rank agree/disagree answers as in §5.1, and for demographics, we treat each answer choice as a binary variable (1 if chosen by a respondent, 0 otherwise). We exclude demographic values for which we have fewer than 5 responses.

Membership in demographic groups does not strongly correlate with answers to any questions in the survey. The strongest correlation is  $\rho_s = -0.25$ , much smaller than the top 10 correlation coefficients between pairs of questions (Table 1). This suggests that there is a diversity of viewpoints

in each demographic category, with more variation within demographics than between demographics.

Nonetheless, there were a few demographics that were particularly predictive of responses. For example, men and women answered many questions differently. Among our respondents, women are more likely to believe in the importance of linguistic theory and inductive bias, believe that language models don’t understand language, and believe that the carbon footprint of training large models should be a concern for NLP researchers. Whereas, men are more likely to believe that the future impact of NLP research will be good, and under-represented minorities are more likely to agree that AI could have catastrophic consequences.

## 5.3 Clustering

To analyze the results beyond pairwise correlations, we identify clusters of opinions with principal component analysis (PCA). To do this, we linearize the agree/disagree questions along  $[-1, 1]$ , with a 0.5 difference between each of **DISAGREE**, **WEAKLY**



DISAGREE, OTHER, WEAKLY AGREE, and AGREE. For demographics, we treat every answer choice as a binary variable. This process gives us a total of 101 features for each respondent.

We run PCA using `scikit-learn` with 34 components, enough to explain 80% of the variance in the data. The variance in the data is fairly long-tailed, with the first 8 components covering 41.1% of the total variance and the remaining 26 explaining 38.9% (with less than 3% of variance explained by each component in the tail). This indicates that perspectives among NLP researchers may be difficult to reduce to a small number of opposing camps (e.g., pro-scale and anti-scale) without missing a great deal of internal disagreement within those groups.

The top four principal components are shown in Table 3. The most prominent cluster of views in the data corresponds to the belief that we should (or shouldn't) be prioritizing large-scale modeling ("Scaling Maximalism"), aligning with the "LM optimist" perspective proposed in §5.1. Another prominent theme is concern with the pace of progress, characterized by a belief that we are making steps towards AGI and that it is an important concern for NLP researchers. While other themes seem to appear in the components after that, no individual cluster of beliefs explains a large amount of the variance in the data, so these clusters are probably not very useful for directly reasoning about the beliefs of individuals, who are combinations of all principal components.

## 6 Discussion

With the NLP Community Metasurvey, we have put concrete numbers to many contentious issues in NLP: the necessity of expert-designed inductive bias, the importance of AGI, whether language models understand language, and more. Perhaps more interestingly, we have also made concrete numbers of the community's *impressions* of these controversies, in some cases confirming what we already believe and in others producing surprises. For example, the idea that mere scale will solve most of NLP is much less controversial (and much less believed) than it is thought to be, and NLP researchers unexpectedly agree that we should do more to incorporate insights and methods from domain sciences, and that we should prioritize problem formulation and task design. Interestingly, very few of the issues we ask about (only the necessity of

linguistic structure and expert-designed inductive bias, Q2-2 and Q2-3) are noticeably more controversial than respondents expected them to be. This could be due to biases from the amplification of controversy (e.g., in social media), or it could just reflect mundane biases in respondent predictions, e.g., regression towards the middle of the range (~50%) under uncertainty.

There are other biases to keep in mind when interpreting our results: the United States is over-represented, and senior researchers and academics probably are as well (§3, Appendix C), not to mention unmeasured population biases based in the personal networks of the authors. Many of our questions have multiple possible interpretations, many rely on vague terms like "plausible" or "major concern," and some rely on comparisons to reference points such as the Industrial Revolution (Q3-3) or "specific, non-trivial results from... linguistics or cognitive science" (Q2-4) which may carry different implications for different readers. Given these issues, it is probably best to view the answers to the questions on this survey as reflecting *something between objective beliefs and signaling behavior*. Agreement or disagreement with a particular statement may indicate where a respondent believes they stand relative to "received wisdom," which would determine what statements would be worth asserting in the context of the status quo; or, a response could be driven by identification with (or rejection of) an already-known ideological camp that the statement is taken to refer to. While these issues affect the way we should interpret the absolute numbers in our results, they should apply equally to the meta-questions, so we believe it is meaningful to compare the survey's actual and predicted results as a way of discovering false sociological beliefs.

We hope the results of the NLP Community Metasurvey can help us update our sociological beliefs to closer match reality, creating common ground for fruitful and productive discourse among NLP researchers as we confront these issues in the course of our work.

### Acknowledgments

We thank the 480 respondents who completed the survey, as well as (especially) our 26 pilot testers. We also thank Markus Anderljung, Noemi Drekler, Amandalynne Paullada, and Lucy Lu Wang for helping ideate and providing early feedback on the survey, John Thickstun for helpful discussions

about the survey’s results, and the anonymous reviewers for useful feedback about its limitations.

This project has benefited from financial support to SB by Eric and Wendy Schmidt (made by recommendation of the Schmidt Futures program) and Apple. This material is based upon work supported by the National Science Foundation under Grant Nos. 1746891, 1922658 and 2046556. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

## Limitations

While we discuss limitations to our survey design and results in §3 and §6, we expand on the issues here.

### Survey Sample and Response Bias

**Sample size and confidence intervals** Our 327 respondents constitute about 5.17% of the total population of our target demographic group. This sample size gives us enough statistical power to meaningfully analyze correlations between questions and demographics (§5), but it leaves a lot of room for potentially strong biases in the results, as we only covered a small portion of the survey population and we did not have a uniform sample.

Furthermore, the confidence intervals on the mean meta-question responses (depicted in black in the visualizations in Appendix D) should be taken with a grain of salt, as they are computed on the basis of the midpoints of the buckets chosen by each respondent (i.e., an answer of 0–20% is treated as 10%), potentially underestimating the variance of respondents’ actual underlying views.

**Twitter** As described in §3, we advertised the survey on Twitter, and >75% of respondents use Twitter to engage with NLP content. This suggests a strong demographic bias towards Twitter users, which may not be representative of the ACL community as a whole. However, even if we are measuring something closer to the public discourse on Twitter, to the extent that comprises a large portion of the public discourse on research, our results may still be useful.

We did advertise in other ways (described in Appendix A), including an email to the ACL Member Portal and sending individualized emails to highly-publishing authors, and a majority of respondents (63%) heard of the survey through something other

than Twitter (Figure 2). However, even through these platforms it’s likely that responses were biased towards the authors’ personal networks, or towards people with personalities or attitudes which would lead them to be more likely to take such a survey.

**Geographic biases** One of the most concerning sources of response bias is geographic. In comparison to official ACL membership statistics, residents of Asian countries are deeply underrepresented, especially China (3% < 9%), India (1% < 5%), and Japan (0.8% < 5%), compared to the United States, which is overrepresented (58% > 35%).

We noticed this disparity, particularly the lack of responses from China, during the survey period. We tried to make extra overtures—as mentioned in Appendix A.3, one of the authors shared the survey on WeChat (and it was reposted by a few other China-based ML researchers), and we individually emailed many of the most highly-published authors in our list (which included many authors based in Asia) as a way of making sure to cover this part of the community. Unfortunately, this yielded few responses from Asia. It’s worth noting that Zhao et al. (2022) describe a significant disconnect between the US and Chinese ML communities (at least in terms of their citation networks). The ACL community may have a similar pattern worth learning more about.

**Industry and career** As noted in §3, our responses are overwhelmingly from academics and skew senior. While we don’t have statistics on the true proportion of academics or junior versus senior authors in our target demographic, it seems likely that this is a source of response bias. This bias is especially relevant to consider for the questions and meta-questions about academia versus industry (Figure 3a, Appendix D.1). In particular, it seems plausible that some of the disparity between answers and predictions, e.g., in Q1-1 (on whether private firms have too much influence) may have arisen from people assuming that many of the other respondents would be from industry. Similar concerns may apply to, e.g., Q2-1 on scaling maximalism and other questions about scaling, which is strongly associated with industry. Unfortunately, we don’t know what the respondents thought about the survey demographics, as we did not include meta-questions in the demographics section.

**Unobservable confounds and statistical correction** The biases above are sources of concern when trying to draw conclusions about the ACL community as a whole. It stands to reason that we may wish to statistically correct for these biases, but we choose not to, for the following reasons:

- We don't have ground-truth data on any of the above demographic characteristics of our target population (actively publishing ACL authors) and can only assess under- or overrepresentation using proxies such as ACL membership. Using these proxies to do statistical correction would introduce another potential source of bias.
- We don't have enough data to provide meaningful signal in some of the underrepresented groups (e.g., residents of Asia), so correcting for these biases would mostly be amplifying noise.
- Adjusting for observed confounds could potentially amplify unobserved confounds—i.e., the small number of Asian residents who did respond are likely not representative of their geographic group at large. This is true to varying degrees for the other demographic variables as well, and would make the results much harder to interpret.

So instead of doing corrections, we prefer to report the results as simply as possible while being up front about potential biases so the reader can draw their own conclusions.

## Questions and Answers

**Strongly worded questions** Many of our questions asked for respondents' views on strongly worded, opinionated statements. This could have biased the survey responses, for example due to framing effects or false presuppositions included in the statement. While we think this is a serious and worthwhile concern, there are several reasons we chose this format anyway:

- The issue is, to some extent, unavoidable, as the point of our survey is to address controversial issues, which by their nature will hinge on strong opinions and contestable presuppositions.
- This approach allows us to format our survey mostly as agree/disagree questions, which

allows for a straightforward meta-question format and uniform computational analysis.

- One method of handling this issue in surveys is to have multiple variants of each question (e.g., both negated and non-negated versions), and show each respondent a random one. However, this would introduce more variables with respect to the wording of the questions, complicating our analysis and creating more difficulties in interpreting the results.
- The valence of opinions (e.g., expressing the view that scaling up *will* versus *won't* solve practical problems) actually does matter for our purposes: as discussed in §6, one way of interpreting what we are doing with these questions is pointing to salient existing ideological camps. The sets of people who would attack or defend an opinion may not be the exact complements of the sets who would, respectively, attack or defend its negation.

Given that we chose this format, we took several measures to try to minimize the influence of potentially leading questions:

- When piloting the study, we asked pilot participants for their views on whether the questions were leading, whether they incorporated unnecessary assumptions, or whether they felt underspecified or tricky to answer, and changed the wording accordingly. We also piloted with groups of varying views (e.g., some groups worked more on AI ethics, others more on machine learning).
- We included several ways for respondents to opt out of questions which they rejected, allowing them to answer INSUFFICIENTLY INFORMED ON THE ISSUE, QUESTION IS ILLPOSED, or PREFER NOT TO SAY.
- For issues where we expected the assumptions to be contentious, particularly relating to artificial general intelligence (Figure 3c, Appendix D.3) and language understanding (Figure 3d, Appendix D.4), we explicitly reminded the users to answer according to *their* preferred definitions.
- We included a free-response feedback box at the end of each section where respondents could clarify their views or express dissatisfaction with the survey. We carefully looked

through this feedback when interpreting the results, and we attempt to summarize respondents' concerns in the detailed results in [Appendix D](#).

**Unclear questions** In writing the questions, we attempt to distill complex and controversial issues into single sentences, and we elicit responses on simplified, 4-point scales. Inevitably, respondents' interpretations of some questions are hard to know for sure, and they varied between respondents, making it more difficult to interpret the results. We can get some insight into this using respondents' free-form feedback. Some examples include the following:

- Q6-6 (on predicting psychological characteristics, discussed in [Appendix D.6](#)) seems to lump together too many disparate issues for us to be able to draw clear conclusions.
- Feedback on Q3-4 (on catastrophic risks of AI) showed that respondents may have answered the question on the basis of various different kinds of scenarios—some said AI causing nuclear war was a ridiculous proposition because it would never be handed the “big red button”, while others felt that a mistake by an AI acting even as a minor indicator could plausibly have catastrophic consequences in a tense enough geopolitical landscape (see [Appendix D.3](#)).

More details are provided for each respective question in [Appendix D](#). Some other potential issues include the following:

- Some of our wording was vague: e.g., the words “plausibly” and “plausible” in Q3-3 and Q3-4 may be read weakly to mean *possible*, or more strongly to mean *likely*. The difference between these two interpretations is great, especially when it comes to high-stakes issues like revolutionary societal change or global catastrophic risk. Likewise, extreme reference points like the Industrial Revolution (Q3-3) or all-out nuclear war (Q3-4) may be too far outside of respondents' typical reference frames for them to provide a well-calibrated answer.
- Our questions point at active controversies and complex issues, and respondents likely answered survey questions fairly quickly. In this

context, respondents may have used attribute substitution heuristics ([Kahneman and Frederick, 2005](#)), wherein they substitute a hard question (e.g., *will AI change society more than the Industrial Revolution?*) for an easier one (e.g., *will AI make things very weird and different pretty soon?*) before answering.

**Interpreting WEAK answers** Our main aim with the [DISAGREE, WEAKLY DISAGREE, WEAKLY AGREE, AGREE] Likert scale is to get respondents to indicate which side of each issue they stand on. To do this, we encourage respondents to choose one of the WEAKLY answers even if they are basically on the fence. For example, “depends, leaning negative” would be an answer of WEAKLY DISAGREE (see [Appendix E](#)). Our approach seems to work for our purpose, as the proportion of OTHER answers is low (<20%) for all questions. However, this means our analysis likely includes a lot of people who do not strongly hold the opinion indicated by their answer. This may be easy to miss in our analyses and meta-survey comparison, which generally group the DISAGREE and WEAKLY DISAGREE answers (and respectively for AGREE) together to binarize the results. So we encourage readers to bear in mind the distinction of WEAK responses when reading the results for a holistic impression of the NLP community.

**Responses are not necessarily indicative of the truth** As stated in §1, the results of our survey are *descriptive* of the NLP community, and the issues we ask about (many of them normative questions) cannot be resolved by majority vote. However, some questions do ask respondents to make (somewhat) falsifiable predictions, e.g., about the state of jobs and publication in NLP ([Appendix D.1](#); Q1-2–Q1-4), the promise of scale, linguistic structure, and inductive bias ([Appendix D.2](#)), and the societal effects of future AI systems ([Appendix D.3](#); Q3-3, Q3-4). But it is still unclear whether aggregating the opinions of NLP researchers on such issues constitutes a reliable forecast of future events.

Uniquely, our methodology does allow us to measure which answers are more popular than expected, corresponding to the *surprisingly popular* answer selection rule for crowd wisdom proposed by [Prelec et al. \(2017\)](#), which better identifies correct answers than majority vote in the case of factual questions where the answer is only known by a well-informed minority. The results of applying

this method to predictions of the future are positive, though more mixed (Lee et al., 2018; Rutchick et al., 2020). While we cannot take the “surprisingly popular” answers as indicative of truth in our case, it is worth considering carefully why some responses were more or less popular than expected.

## References

- Blaise Agüera y Arcas, Margaret Mitchell, and Alexander Todorov. 2017. [Physiognomy’s new clothes](#).
- Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. [Concrete problems in ai safety](#).
- Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. [Palm 2 technical report](#).
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. [Constitutional ai: Harmlessness from ai feedback](#).
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. [On the dangers of stochastic parrots: Can language models be too big?](#) In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’21, page 610–623, New York, NY, USA. Association for Computing Machinery.
- Emily M. Bender and Alexander Koller. 2020. [Climbing towards NLU: On meaning, form, and understanding in the age of data](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.
- Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khat-tab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avani Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R’e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang,

- Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. [On the opportunities and risks of foundation models](#). *ArXiv*.
- Nick Bostrom. 2014. *Superintelligence: Paths, Dangers, Strategies*, 1st edition. Oxford University Press, Inc., USA.
- David Bourget and David J. Chalmers. 2014. [What do philosophers believe?](#) *Philosophical Studies*, 170(3):465–500.
- David Bourget and David J. Chalmers. 2020. Philosophers on philosophy: The PhilPapers 2020 survey.
- Samuel R. Bowman and George Dahl. 2021. [What will it take to fix benchmarking in natural language understanding?](#) In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. [Sparks of artificial general intelligence: Early experiments with gpt-4](#).
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pilla, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. [PaLM: Scaling language modeling with pathways](#). *CoRR*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. [Pathologies of neural models make interpretations difficult](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, Brussels, Belgium. Association for Computational Linguistics.
- Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. 2019. [Risks from learned optimization in advanced machine learning systems](#).
- John P. A. Ioannidis. 2005. Why most published research findings are false. *PLoS Medicine*, 2.
- Sarthak Jain and Byron C. Wallace. 2019. [Attention is not Explanation](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.
- Daniel Kahneman and Shane Frederick. 2005. A model of heuristic judgment. In K. Holyoak and B. Morri-son, editors, *The Cambridge Handbook of Thinking and Reasoning*, pages 267–293. Cambridge University Press.
- Andrey Kurenkov. 2020. Lessons from the PULSE model and discussion. *The Gradient*.
- Michael D Lee, Irina Danileiko, and Julie Vi. 2018. Testing the ability of the surprisingly popular method to predict nfl games. *Judgment and Decision Making*, 13(4):322–333.
- William Merrill, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. 2021. [Provable Limitations of Acquiring Meaning from Ungrounded Form: What Will Future Language Models Understand?](#) *Transactions of the Association for Computational Linguistics*, 9:1047–1060.
- Julian Michael. 2020. [To dissect an octopus: Making sense of the form/meaning debate](#). *Blog post*.
- Sharan Narang, Hyung Won Chung, Yi Tay, Liam Fedus, Thibault Fevry, Michael Matena, Karishma Malkan, Noah Fiedel, Noam Shazeer, Zhenzhong Lan, Yanqi Zhou, Wei Li, Nan Ding, Jake Marcus, Adam Roberts, and Colin Raffel. 2021. [Do transformer modifications transfer across implementations and applications?](#) In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5758–5773, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

OpenAI. 2023. [Gpt-4 technical report](#).

David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluís-Miquel Munguia, Daniel Rothchild, David R. So, Maud Texier, and Jeff Dean. 2022. [The carbon footprint of machine learning training will plateau, then shrink](#). *Computer*, 55(7):18–28.

Christopher Potts. 2020. [Is it possible for language models to achieve language understanding](#). *Medium*.

Dražen Prelec, H. Sebastian Seung, and John McCoy. 2017. [A solution to the single-question crowd wisdom problem](#). *Nature*, 541(7638):532–535.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Inioluwa Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. [AI and the everything in the whole wide world benchmark](#). In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Abraham M Rutchick, Bryan J Ross, Dustin P Calvillo, and Catherine C Mesick. 2020. Does the “surprisingly popular” method yield accurate crowdsourced predictions? *Cognitive research: principles and implications*, 5:1–10.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020. [Green AI](#). *Communications of the ACM*, 63(12):54–63.

Rich Sutton. 2019. [The bitter lesson](#).

Yi Tay, Mostafa Dehghani, Samira Abnar, Hyung Won Chung, William Fedus, Jinfeng Rao, Sharan Narang, Vinh Q Tran, Dani Yogatama, and Donald Metzler. 2022. Scaling laws vs model architectures: How does inductive bias influence scaling? *arXiv preprint arXiv:2207.10551*.

Sarah Wiegrefe and Yuval Pinter. 2019. [Attention is not not explanation](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Bingchen Zhao, Yuling Gu, Jessica Zosa Forde, and Naomi Saphra. 2022. One Venue, Two Conferences: The Separation of Chinese and American Citation Networks. In *NeurIPS Workshop on Cultures in AI*.

## A Methodological Details

### A.1 Choosing Questions

We aimed to ask about issues:

- which are frequently discussed in the community,
- which are the subject of public disagreement,
- about which the NLP community often reflects back on itself, especially where people seem to perceive themselves as in the minority (hot takes) or majority (taking something for granted), and
- for which, if we understand the community’s opinions and meta-opinions better, it may aid our ability to communicate and help people understand how to most effectively communicate about their research.

With these criteria in mind, we (the authors) brainstormed a large initial list of potential questions. After discussing, we voted on which ones to include in the survey, chose roughly the top 30 questions, finalized the agree/disagree question format, and began pilot testing (described in [Appendix A.2](#)), which we used to refine the set of questions, their phrasing, and their presentation format in the survey. The questions used in the survey are shown in = Figures 3 and 4.

### A.2 Pilot Testing

The first author conducted 6 pilot studies with about 26 different participants from Computer Science and Linguistics departments, mostly based in the United States, during the months of February and March of 2022. After pilot participants took the survey, they were asked for feedback in a group Zoom call. Participants were asked about any questions they perceived as leading, reasons they might have refused to answer questions, reasons they might have wanted to stop taking the survey in the middle, and whether the purpose of the survey was clear, etc. The survey instructions and question wording were updated in accordance with their feedback.

### A.3 Distribution and Advertisement

To reach a broad audience of NLP researchers, we set up a homepage for the survey and advertised in the following ways: (a) ACL Member Portal: we sent a call for participation to the ACL membership mailing list. The email included the details of the survey, its purpose, and the charitable donation incentive. (b) ACL 2022 in Dublin: Four of our team members advertised the survey to conference attendees in-person. They distributed flyers/posters of our survey and free stickers that said

“NLP survey” or “I took the NLP survey.” (c) Twitter: We released multiple tweets as advertisement, with the original being retweeted 100+ times. (d) Slack channels: We posted about the survey in the Slack channels of a few labs, as well as an NLP Slack channel with more than 470 members that was set up during ACL 2020. (e) Emails: We attempted to encourage more participants from senior authors by sending personal invitations to 568 authors that have published at least eight qualifying papers since 2019 (we did not exhaustively email all of them, as it required manually sourcing email addresses based on names in the ACL Anthology). (f) Other social media (including WeChat, to encourage participation from researchers in China) and personal interactions with NLP researchers.

#### A.4 Donations

Based on respondent preferences, we donated \$950 to the WHO COVID-19 Solidarity Response Fund,<sup>7</sup> \$1,650 to GiveWell’s Maximum Impact Fund,<sup>8</sup> \$830 to GiveDirectly,<sup>9</sup> and \$1,140 to the Distributed AI Research Institute.<sup>10</sup> 23 respondents (5%) did not provide an answer to this question, so we did not make donations on their behalf.

#### A.5 Data Postprocessing

On the demographic questions for which we received “other” answers (career stage, gender, and how respondents heard about the survey), we manually assign each such response to the closest available answer choice for the purposes of reporting in §3 and Appendix C.

As mentioned in §4.1, 20% of respondents who answered the meta-question on likely sources of future advances ranked “Other” as the most common answer to the question. We assume these were mistakes, since it is unlikely that people would think a plurality of respondents would reject all of the (fairly broad) provided options. We take this, then, to mean the rankings provided by these respondents were probably reversed, with 5 being the most common and 1 the least common. So we reverse the rankings provided by these respondents for the purposes of analysis. Besides changing the “Other”

<sup>7</sup><https://www.who.int/emergencies/diseases/novel-coronavirus-2019/donate>

<sup>8</sup><https://www.givewell.org/maximum-impact-fund>

<sup>9</sup><https://www.givedirectly.org/>

<sup>10</sup><https://www.dair-institute.org/support>

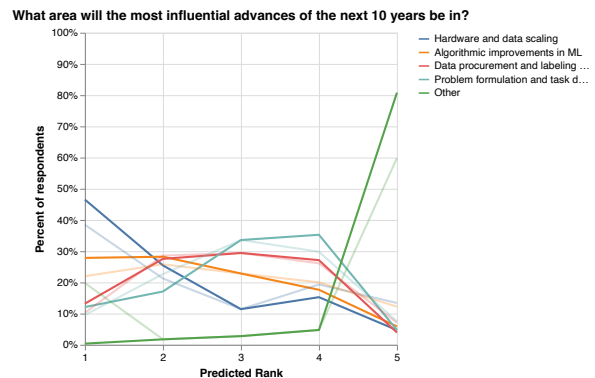


Figure 6: We postprocessed the predicted rankings for likely sources of future advances by reversing the rankings given by all respondents who placed “Other” first (20% of respondents). The rankings for the unadjusted data are shown in the faded lines; besides the change to the “Other” line, overall trends are the same.

statistics, this does not seem to have a noticeable effect on overall trends (Figure 6).

## B Overview of Responses

See Figure 7 for an overview of the responses to agree/disagree questions including OTHER answers. OTHER answers were fairly uncommon—never above 20% for any question—and the most frequent one was INSUFFICIENTLY INFORMED ON THE ISSUE, which many respondents gave for the questions about an NLP winter (Q1-3, Q1-4) and the “wrong track” questions (Q5-{3–6}).

## C Detailed Demographics

Figures 9–12 show full results for demographics questions. Results are restricted to the target demographic of those with at least 2 \*CL publications in the last three years. Numeric labels for percentages below 5% are omitted for space.

## D Detailed Results

In this section we discuss the results of each section of the survey in detail. For each question (for example, in Figure 13), we display the proportion of AGREE, WEAKLY AGREE, WEAKLY DISAGREE, and DISAGREE answers in a band along the bottom of the visualization. These percentages exclude those who gave one of the OTHER answers (INSUFFICIENTLY INFORMED ON THE ISSUE, QUESTION IS ILL-POSED, or PREFER NOT TO SAY), which were relatively rare (<20% of responses for all questions, <10% for 75% of questions; see Appendix B, Figure 7). The vertical



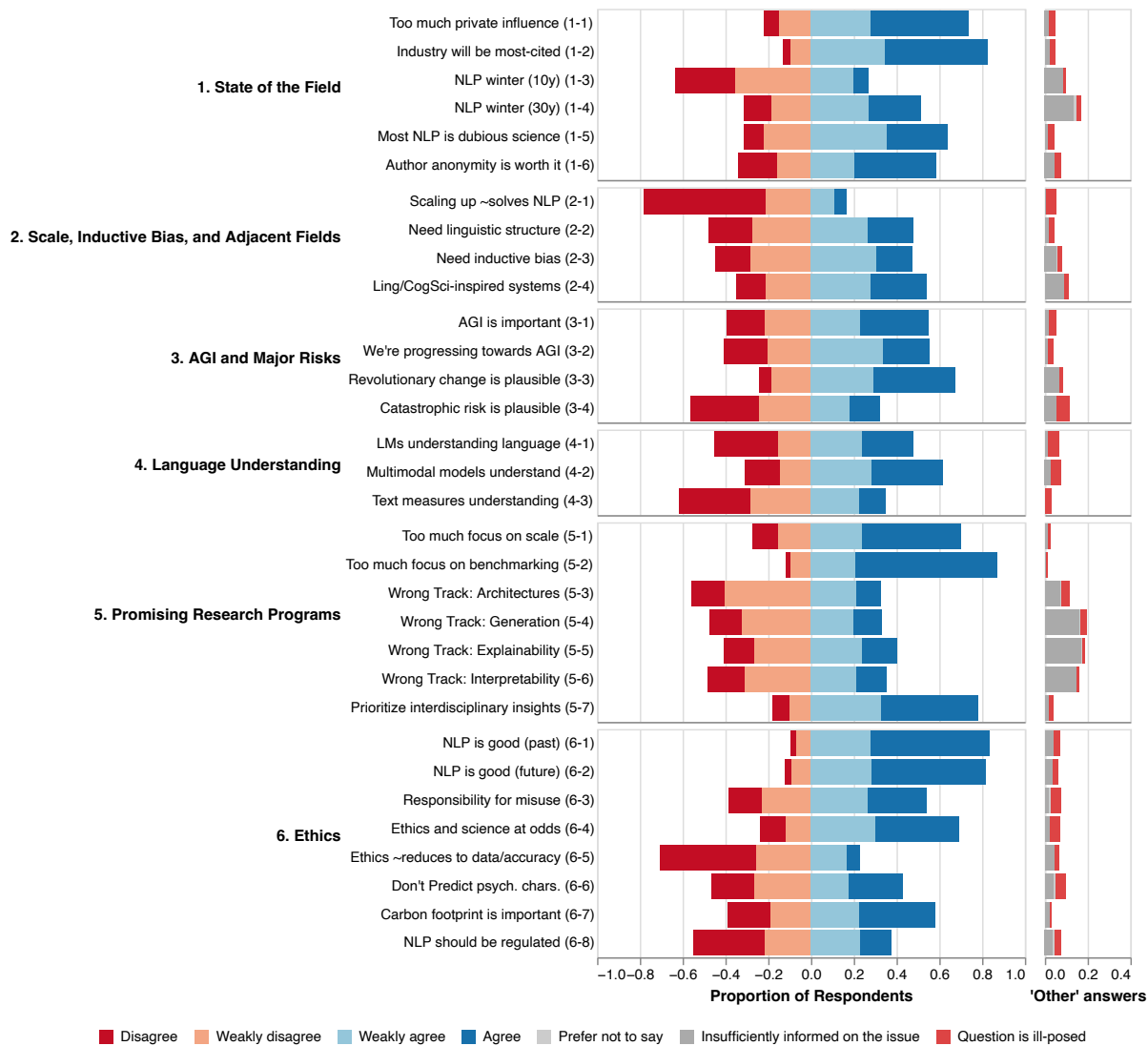


Figure 7: Full overview of responses, including OTHER answers.

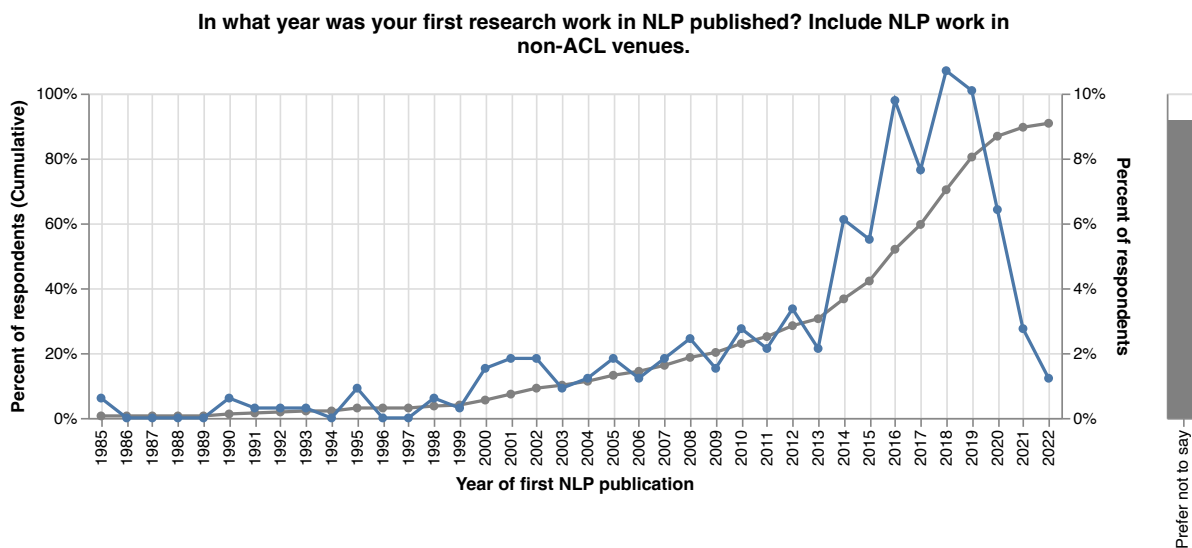


Figure 8: Respondents' year of first NLP publication.

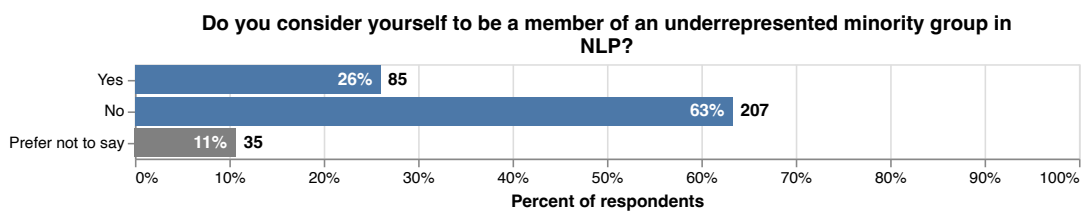
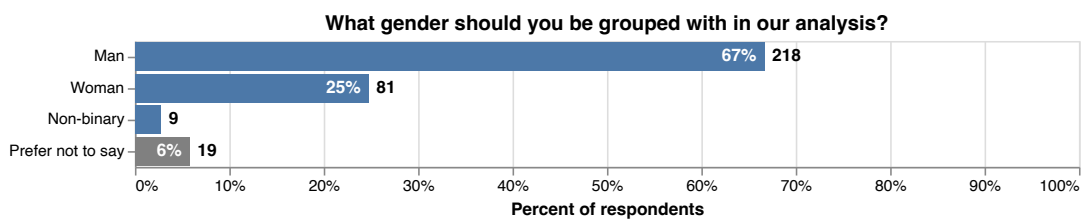
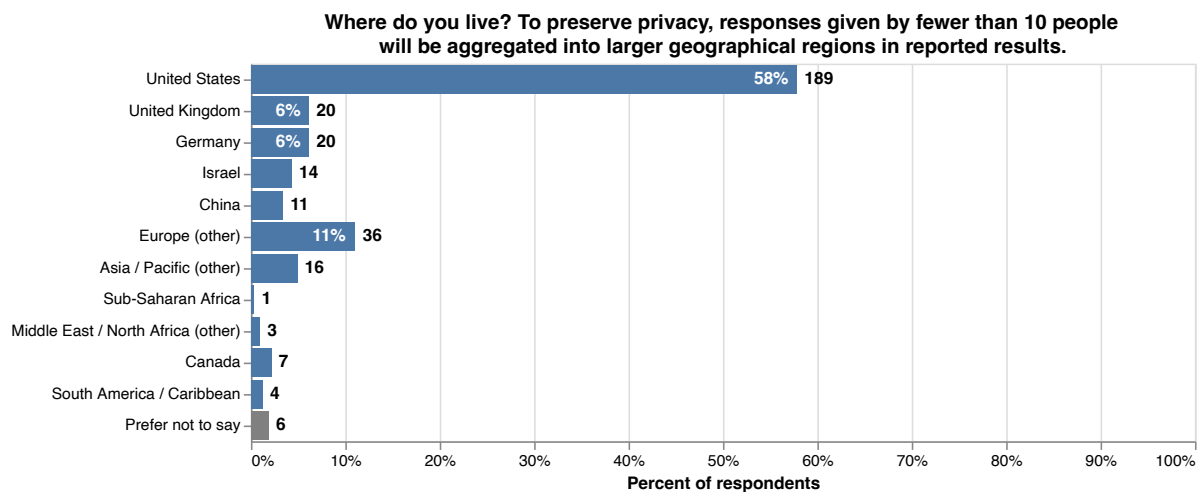


Figure 9: Basic demographics.

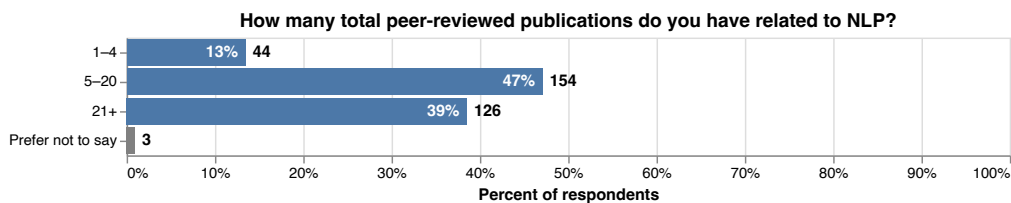
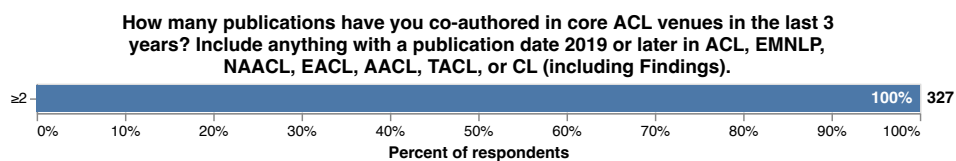
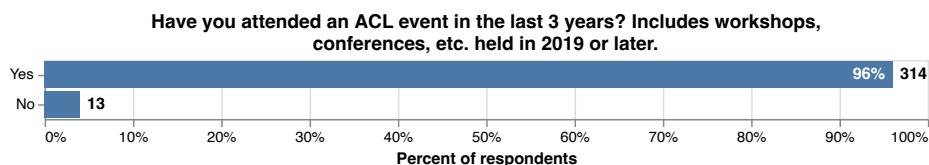
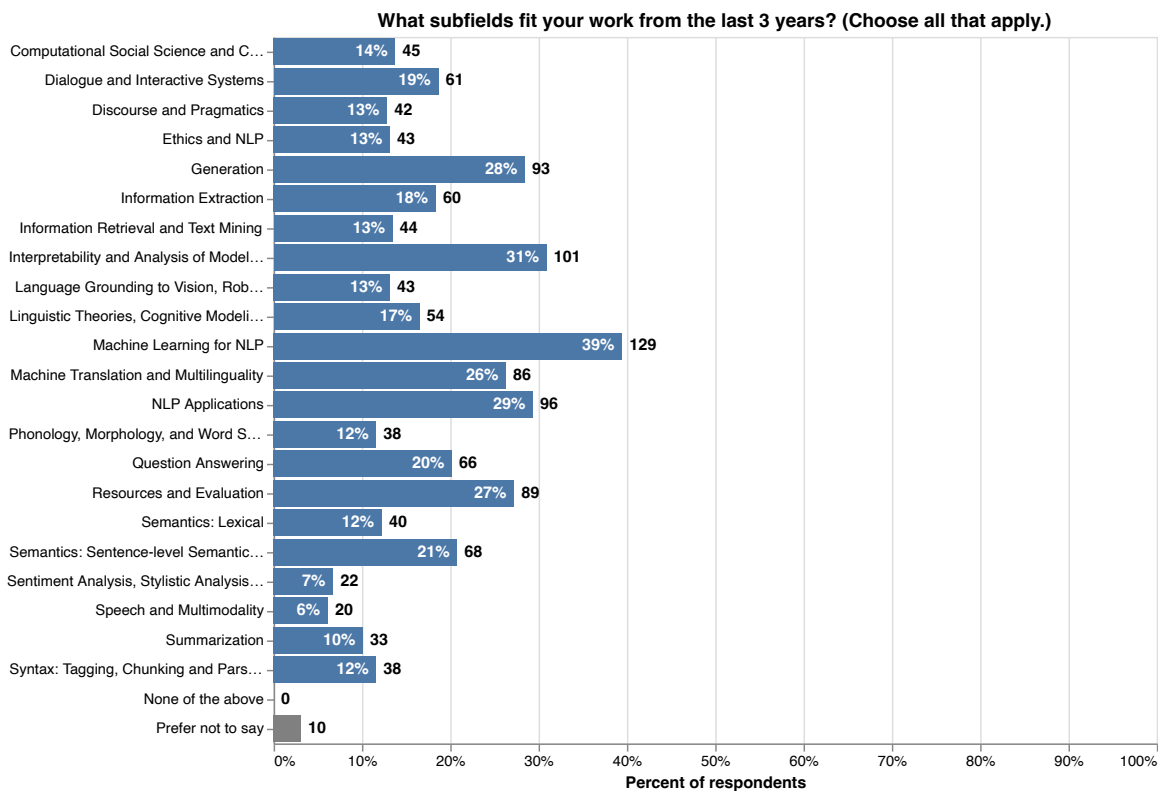


Figure 10: Respondents' research activities. The number of publications in the last 3 years being at least 2 is 100% by construction, since this question is what we used to identify the demographic on which we are reporting results.

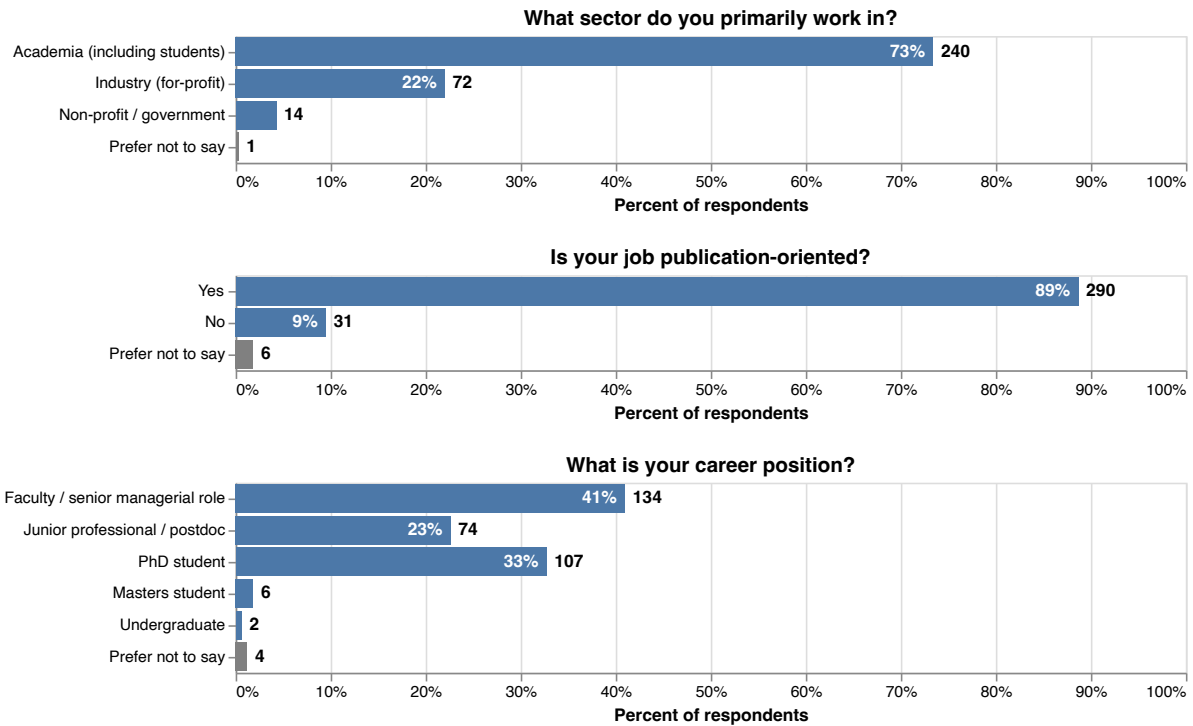


Figure 11: Career demographics.

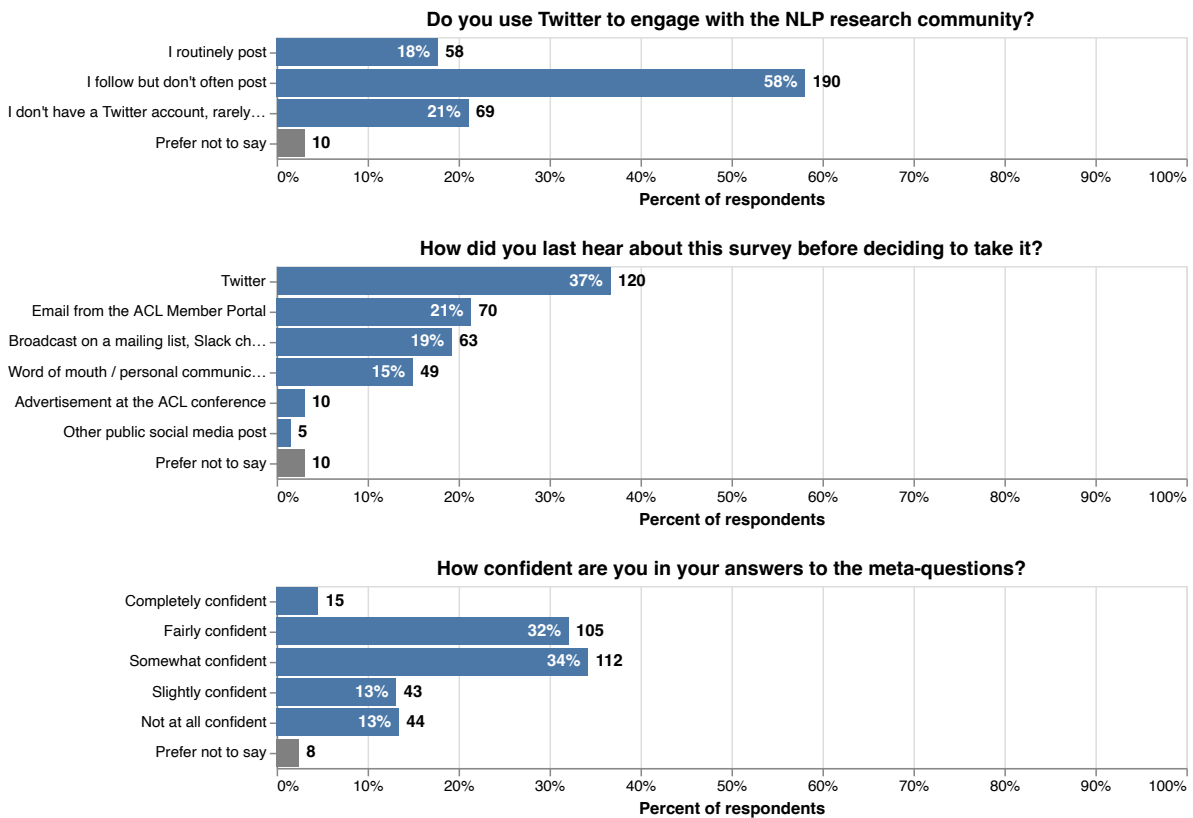


Figure 12: Other information provided by respondents.

green line shows the total percentage who **AGREE** or **WEAKLY AGREE** with the statement, which was what we asked respondents to predict in the meta-questions. The grey bars show the distribution of meta-question answers, each bin aligned with its corresponding range of percentages (0%–20%, 20%–40%, etc.). The green and black dots and bars show the mean and 95% bootstrap confidence intervals of the true and predicted percentage of people who **AGREE** or **WEAKLY AGREE** (treating each meta-question answer bucket as its midpoint).

Unless otherwise stated, all percentages and percentage differences mentioned in this section will be in absolute terms and exclude the ‘other’ answers, and when we refer to respondents “agreeing” with a statement (without special typesetting) we include both **AGREE** and **WEAKLY AGREE** answers (and respectively with **DISAGREE** and **WEAKLY DISAGREE**). In this discussion, we will sometimes break down results by demographic group (e.g., comparing the agreement rates of men and women); unless otherwise stated, all such comparisons correspond to statistically significant differences between the groups ( $p < 0.05$ ) by a bootstrap test. More information, visualizations, and confidence intervals for such comparisons can be found online at <https://nlpsurvey.net/results/>.

## D.1 State of the Field (Figure 13)

The first set of questions asks for opinions about the health of the NLP community.

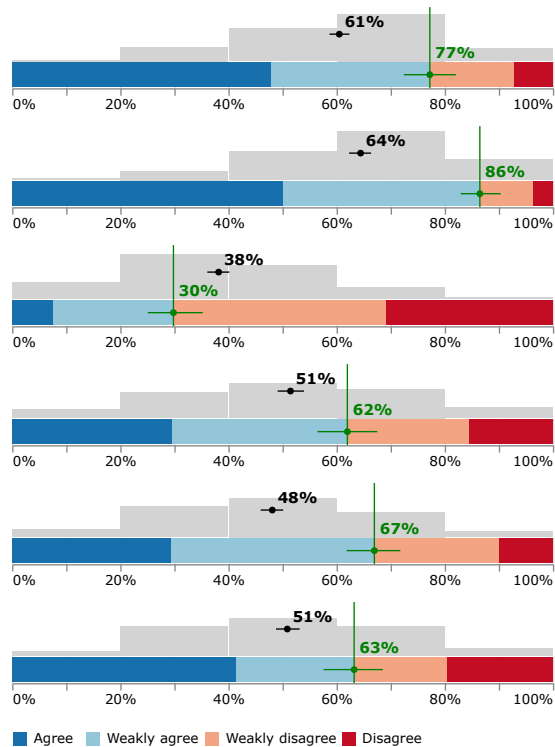
**Industry is seen as having undue influence (Q1-1, Q1-2).** Private firms are overwhelmingly seen as likely to produce the most-cited research of the next 10 years (Q1-2, 82%), but they are also seen as having too much influence (Q1-1, 74%). This suggests, as some respondents pointed out in the survey feedback, that many believe that number of citations is not a good proxy for value or importance. It also suggests a belief that industry’s continued dominance will have a negative effect on the field, perhaps through their singular control of foundational systems such as GPT-3 (Brown et al., 2020) and PaLM (Chowdhery et al., 2022), or from the energy that widely-cited work in pre-training (Devlin et al., 2019; Radford et al., 2019) draws away from other research agendas. Respondents under-predict the popularity of the majority view by more than 15% on both of these questions, suggesting they might believe alternative agendas are already under-prioritized, such as directions fo-

cus on incorporating interdisciplinary insights as opposed to raw scaling, or problem formulation and task design—other under-predicted views, as we will see in Appendix D.2, Appendix D.5, and §4.1).

The under-prediction of agreement on Q1-1 and Q1-2 may also be an artifact of our sample population, which is overwhelmingly academic. Opinions are very different between job sectors, where 82% in academia agree that private firms have too much influence (Q1-1) compared to only 58% of respondents in industry.

**NLP Winter is expected by a majority in the long term (Q1-3, Q1-4).** We ask respondents whether they expect there to be an “NLP winter,” where funding and job opportunities fall by at least 50% from their peak, in the near future. A substantial minority of 30% expect this to happen within the next 10 years (Q1-3), with only 7% **AGREE**ing. For the next 30 years (Q1-4), confidence is much greater, with 62% expecting an NLP winter. Even a minority predicting such a major shift in the field reflects an overall belief that NLP research will undergo substantial changes in the near future (at least, in who is funding it and how much). Further interpretation of these results is difficult: For example, respondents may believe an NLP winter will arrive because the pace of innovation will stall (perhaps the reason they think industry research is overemphasized), because the ability to advance the state of the art will be monopolized by a small number of well-resourced industry labs (as they expect industry to continue producing widely-cited research), or because the distinction between NLP and other AI disciplines will disappear (as suggested by some respondents).

**NLP is mostly seen as a scientifically dubious (Q1-5).** A majority agrees that most NLP work is of “dubious scientific value” (67%). Respondents expressed uncertainty over what should count as “dubious,” as well as concerns about who determines the value of research. On one hand, such research could refer to work which is fundamentally unsound with ill-posed questions and meaningless results, which would be a powerful indictment of NLP research. On the other hand, it could simply mean that many reported findings are of little importance or are not robust, which would arguably not make NLP unique among sciences (Ioannidis, 2005). Either way, this result suggests that many



#### 1-1. Private firms have too much influence

Private firms have too much influence in guiding the trajectory of the field.

#### 1-2. Industry will produce the most widely-cited research

The most widely-cited papers of the next 10 years are more likely to come out of industry than academia.

#### 1-3. NLP winter is coming (10 years)

I expect an "NLP winter" to come within the next 10 years, in which funding and job opportunities in NLP R&D fall by at least 50% from their peak.

#### 1-4. NLP winter is coming (30 years)

I expect an "NLP winter" to come within the next 30 years, in which funding and job opportunities in NLP R&D fall by at least 50% from their peak.

#### 1-5. Most of NLP is dubious science

A majority of the research being published in NLP is of dubious scientific value.

#### 1-6. Author anonymity is worth it

Author anonymity during review is valuable enough to warrant restrictions on the dissemination of research that is under review.

Figure 13: *State of the Field*. Here, and in subsequent such figures, the lower number (in green) represents the fraction of respondents who agree with the position out of all those who took a side. The grey bars show the relative proportion of meta-question predictions in each bin (0–20%, 20–40%, etc.), and the upper number (in black) shows the average predicted rate of agreement, computed treating each bin as its midpoint. The green and black horizontal lines show 95% bootstrap confidence intervals.

NLP researchers think it is worth reflecting deeply on the value of our work. As respondents see the community being less critical than it actually is (by 19% absolute), it might be that those who are critical of the scientific standards of the field are not as likely to voice their views in public, or that vocal critics who exist are seen as less representative of the population than they actually are.

**Anonymity is still controversial (Q1-6).** \*CL conferences have much stricter anonymity policies than many other conferences NLP researchers submit to (e.g., NeurIPS, ICLR, and ICML). Responses suggest the community is in favor of these policies on balance (63% agree anonymity is important enough to warrant restrictions on disseminating preprints), though they are perceived as contentious: respondents guessed that around 51% of the target population would be in favor of such restrictive policies. Since the \*CL anonymity policies have been subject to intense debate on platforms such as Twitter, this suggests that those critical of the policies may have been disproportionately represented in the minds of NLP researchers. This

question was also split by gender, with 77% of women agreeing but only 58% of men—possibly due to concerns or experience with discrimination on the basis of author identity.

## D.2 Scale, Inductive Bias, and Adjacent Fields (Figure 14)

Questions and meta-questions about the long-term potential of scale, inductive bias, and linguistic structure reveal some of the most striking mismatches between respondent attitudes and beliefs about those attitudes. Broadly speaking, the pro-scale and anti-structure views were much less popular than respondents thought they would be.

A common refrain in the era of ever-larger models is the *Bitter Lesson* (Sutton, 2019): “General methods that leverage computation are ultimately the most effective, and by a large margin.” Under this perspective, one may expect linguistic structure or expert-designed inductive biases to be superseded by learning mechanisms operating on fewer, more general principles (given enough training data and model capacity). While the success of deep learning and large language models may be taken

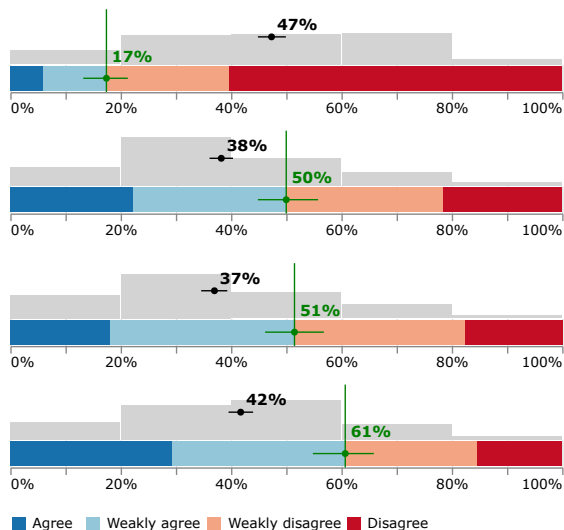


Figure 14: *Scale, Inductive Bias, and Adjacent Fields.*

as supporting evidence for the Bitter Lesson, we find that the community has bought into the Lesson far less than it thinks it has.

**Support for scaling maximalism is greatly over-estimated (Q2-1).** We ask respondents for their views on a strong version of the Bitter Lesson: whether scaling up compute and data resources with established existing techniques can practically solve any important problem in NLP (Q2-1). This is seen as controversial, with respondents predicting a roughly even split of 47% agreement (though variance among predictions was high). However, only a small minority (17%) actually agree with the position, forming the largest discrepancy between predicted and actual opinions in the entire survey. This suggests that the popular discourse around recent developments in scaling up (Chowdhery et al., 2022) may not be reflective of the views of the NLP research community as a whole.

**Trend reversals are predicted for linguistic theory and inductive bias (Q2-2, Q2-3, Q2-4).** The rest of the views articulated in this section were seen as less popular than Q2-1, but in reality they were much more popular (albeit still controversial). On what it will take to practically solve any important problem in NLP, 50% agree that explicit linguistic structure will be necessary (Q2-2), and 51% say the same for expert-designed inductive biases (Q2-3). In addition, 61% of respondents say it’s likely that one of the five most-cited systems in 2030 will take inspiration from clear, non-trivial results from the last 50 years of linguistics or cogni-

### 2-1. Scaling solves practically any important problem

Given resources (i.e., compute and data) that could come to exist this century, scaled-up implementations of established existing techniques will be sufficient to practically solve any important real-world problem or application in NLP.

### 2-2. Linguistic structure is necessary

Discrete general-purpose representations of language structure grounded in linguistic theory (involving, e.g., word sense, syntax, or semantic graphs) will be necessary to practically solve some important real-world problems or applications in NLP.

### 2-3. Expert inductive biases are necessary

Expert-designed strong inductive biases (à la universal grammar, symbolic systems, or cognitively-inspired computational primitives) will be necessary to practically solve some important real-world problems or applications in NLP.

### 2-4. Ling/CogSci will contribute to the most-cited models

It is likely that at least one of the five most-cited systems in 2030 will take clear inspiration from specific, non-trivial results from the last 50 years of research into linguistics or cognitive science.

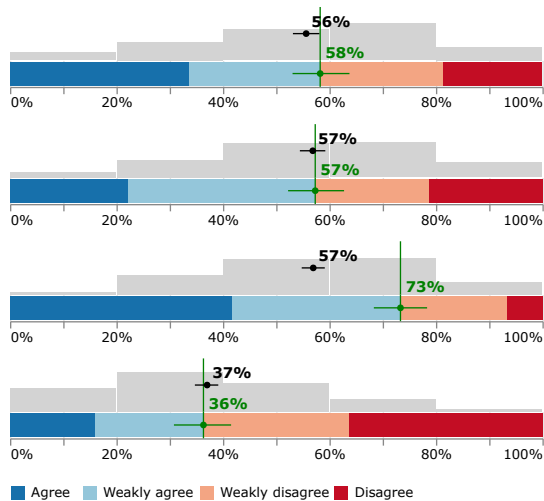
tive science research (Q2-4). All of these views are under-predicted by 12–19%, though the predictions more closely match the responses given by men, as responses to these questions were split by gender. Most notably, women are significantly more likely to agree with Q2-2 that linguistic structure is necessary (65%) compared to men (42%). Women also agreed with Q2-3 (62%) and Q2-4 (69%) more than men (49% and 57%, respectively), but these differences were not statistically significant.

Like many respondents, we find these results surprising. It seems that many believe there will be a reversal of the current trend of end-to-end modeling with low-bias neural network architectures. The results for Q2-4 are particularly surprising to us, as even *today’s* most cited systems seem not to satisfy this requirement, building on little more from cognitive science than a rough construal of *neurons*, *attention*, and *tokens*, which date back much further than 50 years.

## D.3 AGI and Major Risks (Figure 15)

The versatility and impressive language output of large pretrained models such as GPT-3 (Brown et al., 2020) and PaLM (Chowdhery et al., 2022) have prompted renewed discussions about artificial general intelligence (AGI), including predictions of when it might arrive, whether we are actually advancing toward it, and what its consequences would be. In this section, we ask about AGI and some of the largest possible impacts of AI technology.

One concern is that respondents’ answers may depend on their definition of “artificial general



### 3-1. AGI is an important concern

Understanding the potential development of artificial general intelligence (AGI) and the benefits/risks associated with it should be a significant priority for NLP researchers.

### 3-2. Recent progress is moving us towards AGI

Recent developments in large-scale ML modeling (such as in language modeling and reinforcement learning) are significant steps toward the development of AGI.

### 3-3. AI could soon lead to revolutionary societal change

In this century, labor automation caused by advances in AI/ML could plausibly lead to economic restructuring and societal changes on at least the scale of the Industrial Revolution.

### 3-4. AI decisions could cause nuclear-level catastrophe

It is plausible that decisions made by AI or machine learning systems could cause a catastrophe this century that is at least as bad as an all-out nuclear war.

Figure 15: Artificial general intelligence (AGI) and major risks.

intelligence,” and whether they think it is well-defined at all. Our approach to this problem is to deliberately *not* provide a definition (which some respondents would surely find objectionable, no matter which definition we choose). Instead, we instruct respondents to answer according to their preferred definition, i.e., *how they think the community should use the term*, as we view this as an important issue to assess when figuring out how to talk about the issue as a community.

**AGI is a known controversy (Q3-1, Q3-2).** On the questions explicitly about AGI, respondents were mostly split, with 58% agreeing that AGI should be an important concern for NLP researchers (Q3-1) and 57% agreeing that recent research has advanced us toward AGI in a significant way (Q3-2). The two views are highly correlated, with 74% of those who think AGI is important also agreeing with Q3-2 that we’re progressing towards it, while only 37% of people who don’t think AGI is important think we’re making that kind of progress. The meta-responses split similarly to the object-level responses, indicating that the community has a good sense that this is a controversial issue.

It is worth acknowledging what this means: AGI is a controversial issue, the community in aggregate knows that it’s a controversial issue, and now (courtesy of this survey) we can know that we know that it’s controversial. While some may believe that AGI is obviously coming soon, and some may believe that it’s obviously ill-defined, taking either position for granted in the public discourse or scholarly literature may not be an effective way

to communicate to a broad NLP audience; rather, careful and considered discussion of the issue will be more productive for building common ground.

**Revolutionary and catastrophic outcomes are a concern (Q3-3, Q3-4).** 73% of respondents agree that labor automation from AI could plausibly lead to revolutionary societal change in this century, on at least the scale of the Industrial Revolution (Q3-3). This points to a common reason why those who agree with Q3-1 might think AGI is an important concern, especially if we are meaningfully progressing towards it (Q3-2), as it could be fundamentally transformative for society; indeed, all views expressed in this section are positively correlated (see Appendix F). But a significant fraction of respondents (23%) agree with the prospect of revolutionary change (Q3-3) while disagreeing with AGI’s importance, suggesting that discussions about long-term or large-scale impacts of NLP research may not need to be tied up in the AGI debate.

About a third (36%) of respondents agree that it is plausible for AI to cause a major global catastrophe in this century, at least as bad as all-out nuclear war (Q3-4). While this is a much smaller proportion than those expecting revolutionary societal change (Q3-3), the stakes are extremely high and a substantial minority expressing concern about such outcomes indicates that a deep discussion of such risks may be warranted in the NLP community. While we do not ask how specifically respondents think this could happen, potential reasons for such concerns are discussed by Bostrom (2014); Amodei et al. (2016) and Hubinger et al. (2019). Certain



demographics, particularly women (46%) and underrepresented minority groups (53%), were more likely to agree with Q3-4, reflecting pessimism about our ability to manage dangerous future technology perhaps based in the existing track record of disproportionate harm to these groups.

Q3-4 received a lot of critical feedback. Some respondents object to “all-out nuclear war” as far too strong, saying they would agree with less extreme phrasings of the question. This suggests that our result of 36% is an underestimate of respondents who are seriously concerned about negative impacts of AI systems. Some respondents comment that AI/ML systems should not be discussed as if they have agency to make decisions, as all AI “decisions” can be traced back to human decisions regarding training data, architecture, how and on what phenomena models are evaluated (or not), and deployment decisions, among other factors.

#### D.4 Language Understanding (Figure 16)

The question of whether language models understand language has been the subject of some debate in the community (Bender and Koller, 2020; Merrill et al., 2021; Bommasani et al., 2021, §2.6). In this section, we ask some questions relevant to the issue, but one of the challenges is that their answers are highly dependent on how one defines the word “understand.” For this reason, as with Appendix D.3, we deliberately choose *not* to provide a definition, as doing so would risk begging the question or forcing a definition that some would certainly find objectionable. Instead, we instruct respondents to answer according to their preferred definitions, i.e., *how they think the community should use* the word “understand,” as we view this as an important element of the discussion. Many respondents commented that this choice made it harder to respond to the questions in this section, and said they would have preferred a set definition, but only 3–5% responded to any of these questions with QUESTION IS ILL-POSED.

**LMs understanding language is a known controversy (Q4-1, Q4-2).** The question of whether language models can understand language (Q3-1) was split right down the middle, with 51% agreeing. This controversy is reflected in people’s predictions as well, which average to 49% agreement. Many more (67%) agree once the model has access to multimodal data (images, etc.). As with the importance of AGI (Appendix D.3), whether language

models understand language is known to be controversial, and the results of this survey can make it known that it is known. So whatever one’s views are on the issue, it will likely be less useful to take those views for granted as a premise when communicating to a broad NLP audience in the public discourse or scholarly literature. Again, careful and considered discussion of the issue will likely be more productive for building common ground.

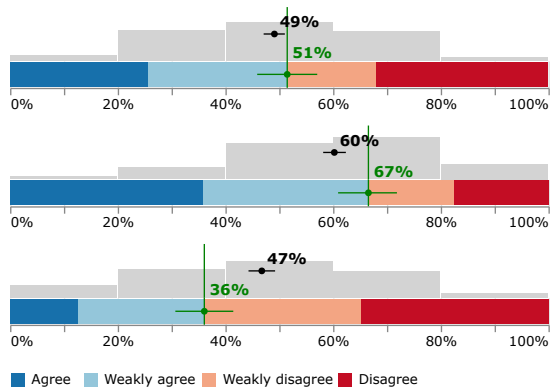
**Understanding may be learnable, but not measurable, using text (Q4-3).** On the question of whether text-only evaluations can measure language understanding (Q4-3), the distribution of predictions was similar to that for language understanding by LMs (Q4-1), averaging 47% predicted agreement. However, unlike Q4-1, only 36% actually agreed with the statement, suggesting that many view it as a separate issue, and some may believe there are things which are learnable from text alone, but cannot be measured using text alone.

Responses to questions in this section vary considerably with respondents’ gender and location. On LMs understanding language (Q4-1), men are more likely to agree (58%) than women (37%), and people in the US are more likely to agree (61%) than those in Europe (31%). There is also a significant gender difference on Q4-3 regarding text-only evaluation of language understanding, where 43% of men agree as opposed to 21% of women.

#### D.5 Promising Research Programs (Figure 17)

In this section, we ask respondents about the kind of research they think the community should be doing, and which research directions they believe are not heading in the right direction. We choose research agendas to ask about based on criticisms, debates, or findings in the literature and public sphere, for example regarding current practice in benchmarking (Bowman and Dahl, 2021; Raji et al., 2021), the relative value of advances in model architectures (Narang et al., 2021; Tay et al., 2022), the use of language models for generation tasks (Bender et al., 2021), and explainability and interpretability of black-box models (Feng et al., 2018; Jain and Wallace, 2019; Wiegrefe and Pinter, 2019).

**Scaling and benchmarking are seen as over-prioritized (Q5-1, Q5-2).** Over 72% of respondents believe that the field focuses too much on scale (Q5-1), a view that was underestimated at 58%. This reflects the same pattern as Q2-1, where



#### 4-1. LMs understand language

Some generative model trained only on text, given enough data and computational resources, could understand natural language in some non-trivial sense.

#### 4-2. Multimodal models understand language

Some multimodal generative model (e.g., one trained with access to images, sensor and actuator data, etc.), given enough data and computational resources, could understand natural language in some non-trivial sense.

#### 4-3. Text-only evaluation can measure language understanding

We can, in principle, evaluate the degree to which a model understands natural language by tracking its performance on text-only classification or language generation benchmarks.

Figure 16: *Language Understanding.*

the prevalence of pro-scale views is overestimated. An even stronger majority of 88% believe there is too much focus on optimizing performance on benchmarks (Q5-2), a view that is highly correlated with Q5-1 (see §5) and is similarly under-predicted at 65%.

**On the wrong track? Opinions vary (Q5-{3-6}).** We ask whether four specific research directions are “on the wrong track”: model architectures (Q5-3), open-ended generation tasks (Q5-4), explainable models (Q5-5), and black-box interpretability (Q5-6). Respondents are divided on these questions, with agreement rates between 37% and 50%, reflecting that these are controversial issues. In most cases, respondents’ predictions also reflect this divide, with a possible exception in explainability (Q5-5), where 50% true agreement is under-predicted at 36%, reflecting that more community members are critical of research in explainable modeling than expected.

While we deliberately used the vague phrase “on the wrong track” to get a sense of people’s general attitudes, some respondents took issue with the framing of these questions; for example, one asks if it means *asking the wrong question* or *finding the wrong solutions*. As such, respondents’ precise interpretations of these questions may vary.

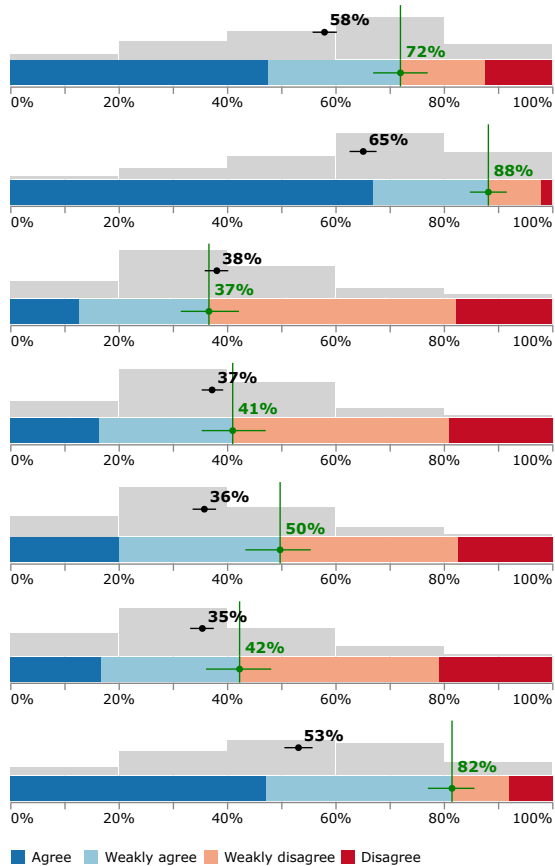
**Interdisciplinary insights are valued more than we think (Q5-7).** The largest disparity between predicted and actual results in this section is on Q5-7, stating that NLP researchers should do more to incorporate insights from relevant domain sciences. While respondents’ predictions about the community’s opinions split this issue down the middle (53%), in reality 82% agree with the view (an outcome only expected by 11% of respondents).

This raises a question: If so many people agree that we should place greater priority on interdisciplinary work (Q5-7), why isn’t more such work already happening? One possible explanation is that the responses to Q5-7 are a form of wishful thinking: Few believe that scale will be sufficient to solve our problems (Q2-1, Q5-1), and many think benchmarks are overemphasized (Q5-2) and insights from sciences like linguistics and cognitive science will be necessary for long-term progress (Q2-2, Q2-3). However, perhaps few know how to actually get results or useful insights from an interdisciplinary approach, leading this kind of work to be underrepresented in the literature and public discourse despite high demand for it. This suggests that the real issue may not be that NLP researchers do not assume interdisciplinary work has anything to offer so much as that we lack the knowledge and tools to make such work effective.

One caveat with this result is that responses vary significantly by job sector; 85% of those in academia agree with Q5-7 compared to 68% of those in industry, and our survey is mostly academics. Despite this difference, even the industry-only agreement rate is underpredicted, so survey response bias likely does not fully explain the mismatch.

## D.6 Ethics (Figure 18)

**NLP is seen as good, and maybe extremely good (Q6-1, Q6-2).** Respondents overwhelmingly regard NLP as having a positive overall impact on the world, both up to the present (89%, Q6-1) and into the future (87%, Q6-2). This strong endorsement of NLP’s future impact stands in contrast with substantial worries about catastrophic outcomes (36%,



#### 5-1. There's too much focus on scale

Currently, the field focuses too much on scaling up machine learning models.

#### 5-2. There's too much focus on benchmarks

Currently, the field focuses too much on optimizing performance on benchmarks.

#### 5-3. On the wrong track: model architectures

The majority of research on model architectures published in the last 5 years is on the wrong track.

#### 5-4. On the wrong track: language generation

The majority of research in open-ended language generation tasks published in the last 5 years is on the wrong track.

#### 5-5. On the wrong track: explainable models

The majority of research in building explainable models published in the last 5 years is on the wrong track.

#### 5-6. On the wrong track: black-box interpretability

The majority of research in interpreting black-box models published in the last 5 years is on the wrong track.

#### 5-7. We should do more to incorporate interdisciplinary insights

Compared to the current state of affairs, NLP researchers should place greater priority on incorporating insights and methods from relevant domain sciences (e.g., sociolinguistics, cognitive science, human-computer interaction).

Figure 17: Promising Research Programs.

Q3-4). While the views are anticorrelated,<sup>11</sup> a substantial minority of 23% of respondents agreed with both Q6-2 and Q3-4, suggesting that they may believe NLP's potential for positive impact is so great that it even outweighs plausible threats to civilization. Whatever this means, it seems clear that many researchers think the stakes of NLP research may be high in the near future. Interestingly, agreement with Q6-1 and Q6-2 are both underpredicted by more than 15%, suggesting that pessimistic voices may be overrepresented in the public discourse.

**Responsibility for misuse: Researchers are somewhat split (Q6-3).** In Q6-3, we ask respondents if they think “it is unethical to build and publicly release a system which can easily be used in harmful ways.” This is admittedly vague, and its answer depends on many factors (e.g., how “easily” the system can be used, how it is released, etc.). Our intent with the question is to get a sense of the

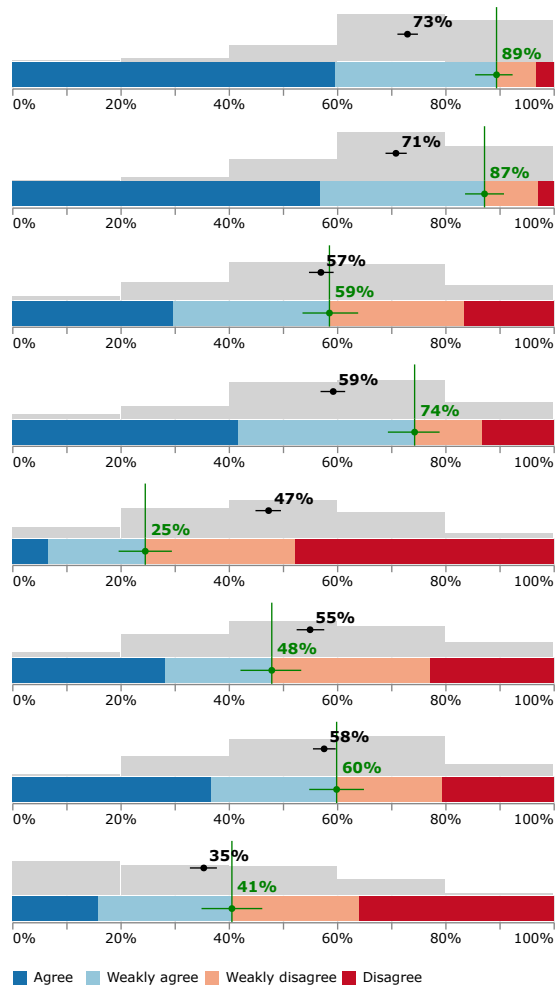
<sup>11</sup>32% of respondents who agreed that NLP will have a positive future impact on society (Q6-2) also agreed that there is a plausible risk of catastrophe (Q3-4), compared to 60% reporting a belief in plausible catastrophic risk among those who disagreed with Q6-2.

degree to which respondents feel that researchers bear ethical responsibility for downstream misuse of the systems that they produce, and assess whether the community's views of itself are accurate in these terms. Responses are somewhat split, with a majority of 59% agreeing, and respondent predictions were reasonably accurate, averaging at 57% predicted agreement. But responses varied by gender, with 74% of women agreeing versus 53% of men. It is worth comparing Q6-3 to Article 1.1 of the ACM Code of Ethics,<sup>12</sup> which is adopted by the ACL<sup>13</sup> and states (among other things): “Computing professionals should consider whether the results of their efforts will... be used in socially responsible ways.”

**Belief in ethical/scientific conflict is underestimated (Q6-4).** When asked if ethical considerations can sometimes be at odds with scientific progress, 74% of respondents agreed—considerably more than the average predicted agree-

<sup>12</sup><https://www.acm.org/code-of-ethics>

<sup>13</sup><https://www.aclweb.org/portal/content/acl-code-ethics>



#### 6-1. NLP's past net impact is good

On net, NLP research has had a positive impact on the world.

#### 6-2. NLP's future net impact is good

On net, NLP research continuing into the future will have a positive impact on the world.

#### 6-3. It is unethical to build easily-misusable systems

It is unethical to build and publicly release a system which can easily be used in harmful ways.

#### 6-4. Ethical and scientific considerations can conflict

In the context of NLP research, ethical considerations can sometimes be at odds with the progress of science.

#### 6-5. Ethical concerns mostly reduce to data quality and model accuracy

The main ethical challenges posed by current ML systems can, in principle, be solved through improvements in data quality/coverage and model accuracy.

#### 6-6. It is unethical to predict psychological characteristics

It is inherently unethical to develop ML systems for predicting people's internal psychological characteristics (e.g., emotions, gender identity, sexual orientation).

#### 6-7. Carbon footprint is a major concern

The carbon footprint of training large models should be a major concern for NLP researchers.

#### 6-8. NLP should be regulated

The development and deployment of NLP systems should be regulated by governments.

Figure 18: *Ethics*.

ment rate of 59%.

There are a couple of potential interpretations of disagreement with Q6-4. On one hand, respondents may believe any ethical problems that come up during the course of NLP research can be solved easily or are trumped by the benefits of scientific progress. On the other hand, they might believe that scientific ‘progress’ which is ethically regressive should not count as ‘progress’ or is inevitably pseudoscientific. Several views in line with the latter (and none with the former) were expressed in the survey feedback, suggesting that it is likely the dominant interpretation among those who disagree with Q6-4. As disagreement was significantly over-predicted by survey respondents, this view may be overrepresented in the public sphere relative to the proportion of NLP researchers who hold it.

**Reduction of ethics to data/accuracy is overestimated (Q6-5).** In light of public debates about the sources and nature of the harms caused by ma-

chine learning systems (Kurenkov, 2020), we ask whether the main ethical challenges posed by current ML systems can be reduced to issues with data quality and model accuracy (Q6-5). It is estimated to be a common view, averaging 47% predicted agreement, but is actually fairly uncommon, with only 25% of respondents agreeing.

**Predicting psychological characteristics is controversial, with caveats (Q6-6).** In light of discussions about surveillance and digital physiology (Aguera y Arcas et al., 2017), we ask whether it is inherently unethical to develop ML systems for predicting internal psychological characteristics like emotions, gender identity, and sexual orientation. Responses were split, with 48% agreeing.

This question received a lot of critical feedback, and it is unclear how much of the split is due to differences in opinion versus interpretations of the question. Some respondents object to grouping transient states (e.g., emotion) with persistent traits

(gender identity, sexual orientation), or say their answer depends on whether the trait is legally protected. Some say it depends on the inputs available to the model, and others say that it may not be *inherently* unethical but is ethically permissible in only a tiny set of carefully considered use cases. Which of these elements of context respondents assumed may have played a major role in determining their answers, and future surveys on these issues might benefit from splitting Q6-6 into several questions.

#### **Carbon footprint is a concern for many (Q6-7).**

A majority of 60% agree with the statement that the carbon footprint of training large models should be a major concern for NLP researchers (Q6-8). This concern is based in part on trends in computation for machine learning at large scale, as Schwartz et al. (2020) note a 300,000x increase in computation over 6 years leading up to 2019. Following this, Patterson et al. (2022) argue that advances in model efficiency and energy management can soon lead to a plateau in energy use from training machine learning models. Both argue that accountability and reporting of energy use is important for keeping the future carbon footprint of training ML models under control. The responses to Q6-8 indicate that a majority of the community would likely appreciate explicit reporting of energy use in NLP publications as well as work that increases the compute efficiency of model training. Responses to this question varied greatly by gender, with 78% of women agreeing as opposed to 51% of men.

#### **NLP researchers are skeptical of regulation (Q6-8).**

Finally, we ask if the development and deployment of NLP systems should be regulated by governments (Q6-8). 41% of respondents agree, and while predictions are accurate on average, a large contingent (31%) of respondents predicted a very low agreement rate of 0–20%. We intend Q6-8 as a weak statement, i.e., that there should be *any* regulations around the development and deployment of NLP systems. However, respondents ask for more nuance, remarking that the answer depends on development versus deployment, details about use cases, and whether we only mean NLP-specific regulations or also include more general regulations on things like energy use or data privacy. As respondents may have come to this question with different assumptions or interpretations around such issues, it is hard to read into the specific implications of this result, except that respondents express a gen-

eral skepticism of government regulation.

## **E Survey Instructions**

The text of our consent form is reproduced in Figure 19 and the survey instructions are reproduced in Figure 20, and examples of how it looks in the browser are given in Figure 21.

## **F Correlation Matrices**

Spearman correlations between questions are shown in Figure 22, and correlations between questions and demographic variables are shown in Figure 23. See §5 for a discussion of the correlation analysis.

### **Consent Form for IRB-FY2022-6461**

You have been invited to take part in a research study to learn more about the beliefs that natural language processing (NLP) researchers hold about the NLP research field, as well as their corresponding meta-beliefs about what other NLP researchers think. This study will be conducted by Prof. Samuel R. Bowman, CIMS - Center for Data Science, Courant Institute of Mathematical Sciences, New York University.

If you agree to be in this study, you will be asked to do the following:

- Complete a questionnaire reporting your beliefs on a variety of potentially controversial issues debated in the NLP community.
- Report some of your own demographic characteristics and general information about your publication and research experience.

Participation in this study will take approximately 20 minutes. There are no known risks associated with your participation in this research beyond those of everyday life.

Although you will receive no direct benefits, the investigators will donate \$10 to a non-profit organization or fund of your choice (four options are listed at the end of the survey) on your behalf, with the total capped at \$10,000 distributed in proportion to the preferences of the first 1,000 people who complete the survey.

Confidentiality of your research records will be strictly maintained by keeping the full data, with opt-in de-anonymization (email addresses only), in a private directory in the cloud accessible only to the investigators. Email addresses will be stripped from this data before sharing with a small group of researchers to analyze, and we will only publicly share statistical aggregates to minimize the risk of de-anonymization after the fact. Information not containing identifiers may be used in future research, shared with other researchers, or placed in a data repository without your additional consent.

Participation in this study is voluntary. You may refuse to participate or withdraw at any time without penalty. You have the right to skip or not answer any questions by selecting a prefer not to say option.

If there is anything about the study or your participation that is unclear or that you do not understand, or if you have questions or wish to report a research-related problem, you may contact Samuel R. Bowman at [bowman@nyu.edu](mailto:bowman@nyu.edu), 60 5th Ave. New York, NY, 10011. For questions about your rights as a research participant, you may contact the University Committee on Activities Involving Human Subjects (UCAIHS), New York University, 665 Broadway, Suite 804, New York, New York, 10012, at [ask.humansubjects@nyu.edu](mailto:ask.humansubjects@nyu.edu) or (212) 998-4808. Please reference the study # (IRB-FY2022-6461) when contacting the IRB (UCAIHS).

You may save a copy of this consent document to keep.

Figure 19: Full text of the survey consent form.

## The NLP Community Metasurvey

This should take ~20 minutes to complete. **For the first 1,000 respondents, we will donate \$10 on your behalf to one of several non-profits that you choose at the end of the survey.**

This is a survey of opinions on issues being publicly discussed in NLP. We (researchers at UW and NYU) invite anyone doing NLP research to take it, though our primary target demographic is **people who have authored or co-authored at least 2 publications in core ACL venues in the past 3 years**. Please share this survey widely — we hope to cover as much of the target demographic as possible.

For each statement, mark whether you **agree** or **disagree**. Then, you will report what percentage of community members you think agree with the statement. This will give a sense of whether our community's impression of itself aligns with its members' actual beliefs, and help us improve this alignment, communicate better, and motivate our work more effectively. More details about our motivation can be found at [nlpsurvey.net](http://nlpsurvey.net). This was inspired by the PhilPapers surveys ([philpapers.org/surveys](http://philpapers.org/surveys)).

**How to Answer** You will be shown a statement and asked where you stand on an **agree/disagree spectrum**:

- Agree
- Weakly agree
- Weakly disagree
- Disagree

Choose the answer that best reflects your views. In our analysis, we will interpret “weakly agree” and “weakly disagree” to include marginal views just barely agreeing or disagreeing (e.g., “depends, leaning positive/negative”). However, in case you cannot place yourself on either side of an issue, there will also be three non-answer options:

- **Insufficiently informed on the issue:** You don't understand the statement or its subject matter well enough to form an opinion.
- **Question is ill-posed:** You reject the distinction between agreeing and disagreeing, or don't think the statement admits any coherent interpretation.
- **Prefer not to say:** You don't feel comfortable providing any of the other answer choices.

If you pick “question is ill-posed” or “prefer not to say,” we would appreciate (optional) feedback at the end of the respective section explaining your reasons so we can better interpret the results.

**Meta-Questions** At the end of each section, you'll be asked to predict what proportion of people on the agree/disagree spectrum will answer **either “agree” or “weakly agree”** to each question.


For the purpose of these questions, please predict relative to the target demographic: people with at least 2 publications in core ACL venues in the last 3 years. For our purposes, core venues are ACL, EMNLP, NAACL, EACL, AACL, TACL, and CL (including Findings). By our count from the ACL Anthology, this includes approximately 5,650 people.

Even if you don't have a strong sense of the community's stance, give your best guess (unless you really feel like you have no priors, in which case you can skip these questions). At the end of the survey, you will rate your overall confidence in the meta-survey questions so we can account for it in our analysis.

**Privacy** Your responses are anonymous and individual responses will not be released publicly. You will have the option to de-anonymize yourself at the end; this will help us audit the results and follow up with you after the survey, but will not be released publicly or shared with anyone without your permission. In accordance with the General Data Protection Regulation (GDPR), all survey respondents have rights over their personally-identifiable information. This form ([nlpsurvey.net/gdpr.pdf](http://nlpsurvey.net/gdpr.pdf)) outlines those rights and how to exercise them. A list of the people who will have access to the non-anonymized data is available at [nlpsurvey.net/about](http://nlpsurvey.net/about).

You can reach us with questions and concerns at [nlp-metasurvey-admin@nyu.edu](mailto:nlp-metasurvey-admin@nyu.edu).

Figure 20: Full text of the survey instructions.

 NEW YORK UNIVERSITY

## The NLP Community Metasurvey

This should take ~20 minutes to complete. **For the first 1,000 respondents, we will donate \$10 on your behalf to one of several non-profits that you choose at the end of the survey.**

This is a survey of opinions on issues being publicly discussed in NLP. We (researchers at UW and NYU) invite anyone doing NLP research to take it, though our primary target demographic is **people who have authored or co-authored at least 2 publications in core ACL venues in the past 3 years.** Please share this survey widely — we hope to cover as much of the target demographic as possible.

For each statement, mark whether you **agree** or **disagree**. Then, you will report what percentage of community members you think agree with the statement. This will give a sense of whether our community's impression of itself aligns with its members' actual beliefs, and help us improve this alignment, communicate better, and motivate our work more effectively. More details about our motivation can be found at [nlp-survey.net](https://nlp-survey.net). This was inspired by the PhilPapers surveys ([philpapers.org/surveys](https://philpapers.org/surveys)).

### How to Answer

You will be shown a statement and asked where you stand on an **agree/disagree spectrum**:

- Agree
- Weakly agree
- Weakly disagree
- Disagree

Choose the answer that best reflects your views. In our analysis, we will interpret "weakly agree" and "weakly disagree" to include marginal views just

Disagree

Other


Insufficiently informed on the issue

Question is ill-posed

Prefer not to say

Of those on the agree/disagree spectrum, what percentage of community members do you think will mark "agree" or "weakly agree" to each statement?  
To ground your judgment, "community members" = people with at least 2 publications in core ACL venues in the last 3 years. You can skip questions where you have no idea, but best guesses are highly encouraged.

	0–20%	20–40%	40–60%	60–80%	80–100%
Private firms have too much influence in guiding the trajectory of the field.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The most widely-cited papers of the next 10 years are more likely to come out of <b>industry</b> than <b>academia</b> .	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I expect an "NLP winter" to come within the next <b>10 years</b> , in which funding and job opportunities in NLP R&D fall by at least 50% from their	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

 NEW YORK UNIVERSITY

## State of the Field

For each of the following statements, mark the answer that best reflects your position. At the end of this section, you will predict what percentage of community members will have marked "agree" or "weakly agree" for each.

Private firms have too much influence in guiding the trajectory of the field.

Agree/Disagree

Agree

Weakly agree

Weakly disagree

Disagree

Other

Insufficiently informed on the issue

Question is ill-posed

Prefer not to say

peak.

I expect an "NLP winter" to come within the next **30 years**, in which funding and job opportunities in NLP R&D fall by at least 50% from their peak.

A majority of the research being published in NLP is of dubious scientific value.

Author anonymity during review is valuable enough to warrant restrictions on the dissemination of research that is under review.

	0–20%	20–40%	40–60%	60–80%	80–100%
I expect an "NLP winter" to come within the next <b>30 years</b> , in which funding and job opportunities in NLP R&D fall by at least 50% from their peak.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
A majority of the research being published in NLP is of dubious scientific value.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Author anonymity during review is valuable enough to warrant restrictions on the dissemination of research that is under review.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

(Optional) Any comments or feedback on this section?

← →

Figure 21: How the survey looks to respondents in a web browser.



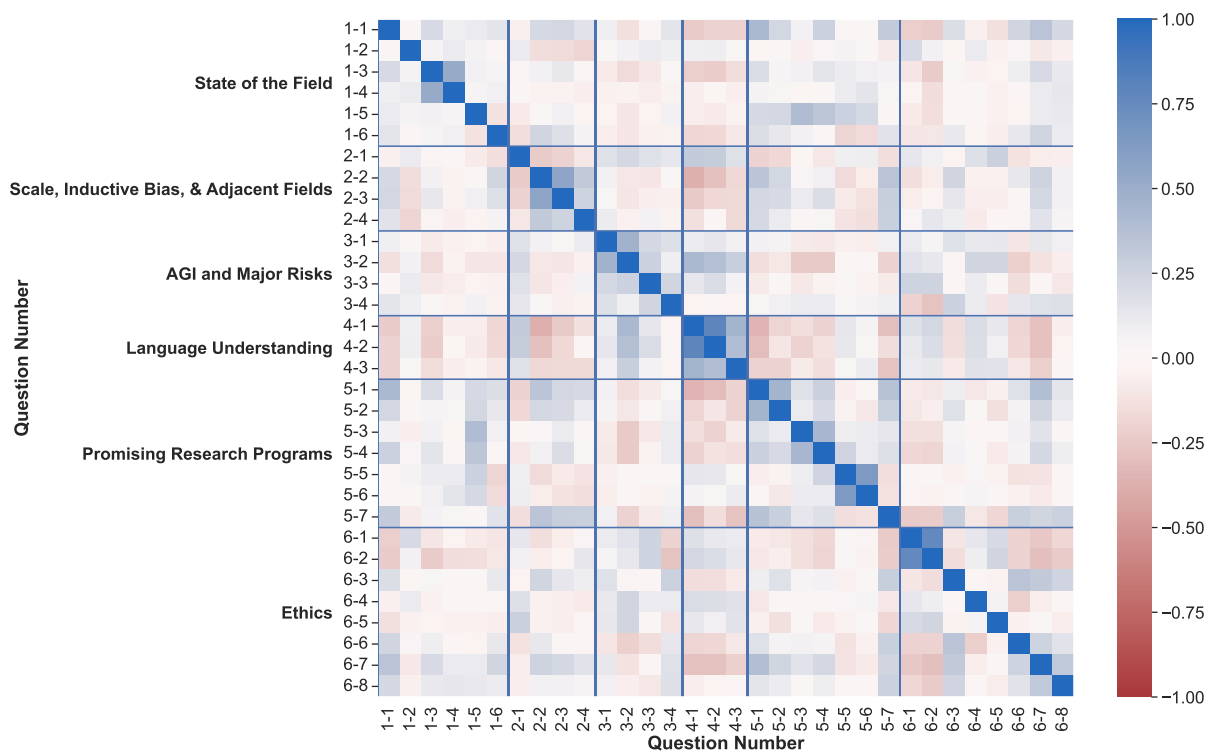


Figure 22: Spearman correlations between answers to all pairs of agree/disagree questions. Lines separate sections of the survey. Question numbers are given in Figures 3 and 4.

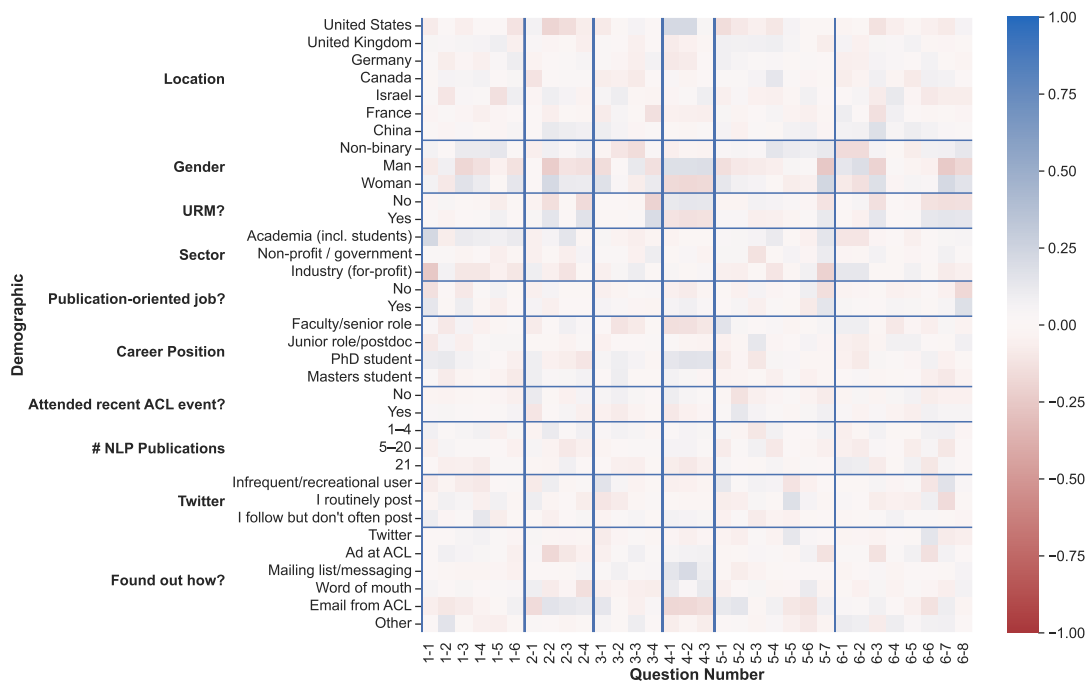


Figure 23: Spearman correlations between membership in demographic groups and answers to agree/disagree questions. Lines separate survey sections and demographic variables. “URM” stands for under-represented minority. We only show demographic values with > 5 respondents. Question numbers are given in Figures 3 and 4.

## ACL 2023 Responsible NLP Checklist

---

### A For every submission:

- A1. Did you describe the limitations of your work?  
*Limitations section, as well as Secs. 3 and 6 (briefly in context).*
- A2. Did you discuss any potential risks of your work?  
*The work was a survey of the NLP community and posed no risks to participants beyond those of everyday life. Our study was approved by our institutional IRB with exempt status.*
- A3. Do the abstract and introduction summarize the paper's main claims?  
*Abstract and Section 1.*
- A4. Have you used AI writing assistants when working on this paper?  
*Left blank.*

### B Did you use or create scientific artifacts?

*Left blank.*

- B1. Did you cite the creators of artifacts you used?  
*Not applicable. Left blank.*
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?  
*Not applicable. Left blank.*
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?  
*Not applicable. Left blank.*
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?  
*Informed consent was obtained for personally-identifiable information. The consent form and details on protection of that information are given in Figure 19 (Appendix E).*
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?  
*Section 3 and Appendix C.*
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.  
*Section 3 and Appendix C.*

### C Did you run computational experiments?

*Left blank.*

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?  
*Not applicable. Left blank.*

---

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

*Not applicable. Left blank.*

- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

*Not applicable. Left blank.*

- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

*Not applicable. Left blank.*

**D  Did you use human annotators (e.g., crowdworkers) or research with human participants?**

*Sections 2 (methodology) and 3 (demographics).*

- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

*Appendix E, Figures 19–21.*

- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?

*Section 2 and Appendices A.3 and A.4.*

- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?

*Section 2 and Appendix E, Figure 19.*

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

*Section 2, 'Platform and Distribution'; Appendix E, Figure 19.*

- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

*Section 3 and Appendix C.*