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

Julian Michael[1,2] **Ari Holtzman**[1] **Alicia Parrish**[4] **Aaron Mueller**[5] **Alex Wang**[3]
Angelica Chen[2] **Divyam Madaan**[3] **Nikita Nangia**[2]
Richard Yuanzhe Pang[3] **Jason Phang**[2] and
Samuel R. Bowman[2,3,4]
[1]Paul G. Allen School for Computer Science & Engineering, University of Washington
[2]Center for Data Science, New York University
[3]Courant Institute of Mathematical Sciences, New York University
[4]Department of Linguistics, New York University
[5]Department of Computer Science, Johns Hopkins University
nlp-metasurvey-admin@nyu.edu

Abstract

We present the results of the NLP Community Metasurvey. Run from May to June 2022, the survey 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 almost exactly in half on questions about 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 not only to gain insight on the spectrum of beliefs held by NLP researchers, but also to uncover *false sociological beliefs* where the community's predictions don't match reality. We find such mismatches on a wide range of issues. Among other results, 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 own 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? Will they ever?
- Is the traditional paradigm of model benchmarking still tenable?
- What kinds of predictive models are ethical for researchers to build and release?
- Will the next most influential advances come from industry or academic labs?

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:

- An early-career researcher may avoid researching a topic they think is important if they think that it is not valued by the community and it would be hard to get past review.

- Extra time and effort may be spent in papers and research agendas defending what is thought to be a provocative position, when it is in fact already well established and widely believed.
- Papers or public communications generally appeal to premises they believe are settled or accepted as general consensus, but if the premise is not actually settled, then the argument will not be convincing to a large portion of the community, and the discourse may become more fractured, with increasingly isolated groups of researchers forming echo chambers around different sets of assumptions.

The NLP research community gets to know itself in a variety of ways: discussions with colleagues at the same institution, presentations and interactions at conferences, invited talks and panel discussions, and interactions on social media like Twitter. All of these sources of information are biased, for example through self-selection towards similar people and amplification of already-prominent or controversial voices. These effects can make it difficult for any individual to get a sense of the NLP research community's beliefs as a whole.

For these reasons, we believe it is worth trying to objectively assess NLP researchers' collective stance on controversial issues. So from May to June 2022, we conducted the **NLP Community Metasurvey**. On each issue, we present a stance, such as *Currently, the field focuses too much on scaling up machine learning models* (Q5-1), and ask respondents to state whether they agree or disagree. Then we ask them to also *predict the percentage 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 & Chalmers, 2014; 2020), a research effort created and maintained by philosophers to assess the philosophy community's beliefs about current topics in their field, along with a separate metasurvey to assess philosophers' sociological beliefs about their professional community. As Bourget & Chalmers write:

> Most of us have had the experience of reading philosophical papers that make sweeping sociological claims about the field that seem quite questionable. It is arguable that the "received wisdom" in a field (roughly, what people in a field take as a default view, one that can sometimes be presupposed) is not based on what most people in the field think, but rather is based on what most people think most people think. If the received wisdom is grounded in false sociological beliefs, that is worth knowing. It occurred to us that it would not be all that difficult to actually gather data about these matters, and that doing so might be an interesting and informative exercise.

The rest of this document reports the methodology and results of the NLP Community Metasurvey. Some key results include:

- A large majority of respondents are against the hypothesis that scaling up current systems and methods could solve "practically any important problem" in NLP, and this view is perceived as much more popular than it is (§4.2,Q2-1). Yet, narrow majorities of respondents also regard recent progress in large-scale modeling as progress towards AGI, and think AGI should concern NLP researchers (§4.3, Q3-1, Q3-2).
- A majority of respondents think that the scientific value of the majority of work in NLP is dubious (§4.1, Q1-5).
- A large majority of respondents believe that NLP researchers should give higher priority to incorporating insights from neighboring fields, and greatly underestimate the number of other NLP researchers that share this belief (§4.5, Q5-7).
- Large majorities of respondents believe NLP research has had a positive impact on the world (§4.6, Q6-1), will have a positive impact in the future (Q6-2), and could plausibly transform society (§4.3, Q3-3). Despite this optimism, a substantial minority also foresee plausible risks of a major global catastrophe caused by ML systems (Q3-4).
- A plurality of respondents think the most influential area of advances in the next 10 years will be in *problem formulation and task design*, as opposed to *hardware and data scaling*, which respondents believed would be the most popular opinion (§4.7).

It is worth noting that 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 the NLP research community.

2 METHODOLOGY

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 C), 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 Figure 2.

Target Demographic: Active Authors in ACL Since we are interested in the public and scientific discourse of the NLP community, we target the survey at active NLP researchers. As an objective criterion, we define the target population of the survey as anyone who is an author on at least two *CL papers published in the last three years. This definition allows us to assess response bias by comparing to ACL members (§3) and provides an objectively defined reference group for survey respondents to make predictions about in the meta-questions.

Question Format We present questions in thematic groupings (e.g., "State of the Field"), phrased as statements that people could indicate their (dis)agreement with. They can respond to each statement on a 4-point scale of AGREE, WEAKLY AGREE, WEAKLY DISAGREE, and DISAGREE. We choose not to include an option in the middle of the scale to indicate a neutral position because our intent is to push respondents to consider where they actually stand. We instruct respondents to choose WEAKLY AGREE or WEAKLY DISAGREE if they have even slight preferences for one side or the other (e.g., "depends, leaning negative"). However, it is sometimes the case that someone truly cannot make a judgment, and for these cases we include three OTHER answers: QUESTION ILL-POSED, INSUFFICIENTLY INFORMED ON THE ISSUE, and PREFER NOT TO SAY.

At the end of each section of the survey, we ask respondents to predict the proportion of respondents in our target population who will either AGREE or WEAKLY AGREE with each statement. Respondents can answer with one of five buckets: 0–20%, 20–40%, 40–60%, 60–80%, or 80–100%. Respondents can also skip the meta-questions, but we encourage them to give their best guess even if they are not sure. Finally, each section has a free-response box for the respondent to provide any comments, criticism, or other feedback on the survey.

The survey instructions are reproduced in full in Appendix D.1.

Platform and Distribution To host the survey, we used NYU Qualtrics. Following the guidelines set out by the NYU Institutional Review Board (protocol FY2022-6461), all respondents gave informed consent before beginning the survey and could either refuse to answer (i.e., by responding PREFER NOT TO SAY) or skip each question.

As an incentive for participation, we committed to donating $10 for each respondent to one of several non-profit organizations that the respondent chooses at the end of the survey.[1]

[1] Our final donations were $950 to the WHO COVID-19 Solidarity Response Fund (https://www.who.int/emergencies/diseases/novel-coronavirus-2019/donate), $1,650 to GiveWell's Max-

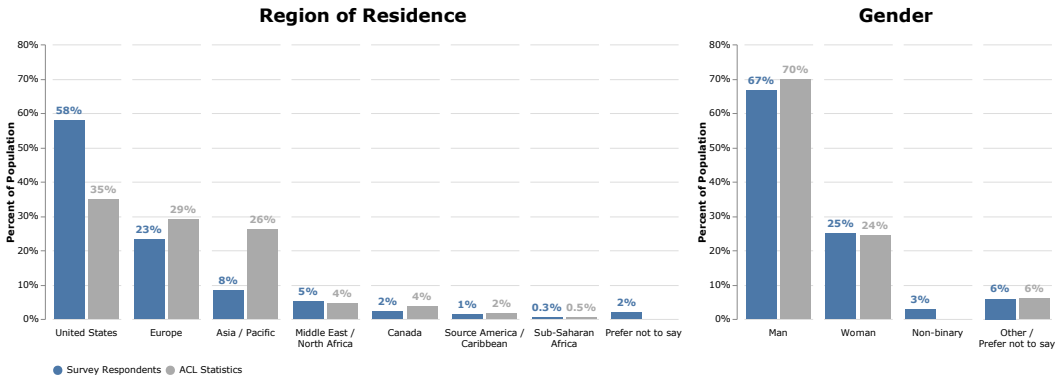


Figure 1: Basic demographics of survey respondents, compared to available statistics from the ACL. For region, we compare to ACL memberships as of summer 2021, and for gender, we use the diversity statistics currently available from the ACL, based on attendance at ACL 2017 in Vancouver (which lacked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

To attempt to reach a broad audience of NLP researchers, we set up a homepage for the survey at https://nlpsurvey.net 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 posting on WeChat to encourage participation from researchers in China) and personal interactions with NLP researchers.

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 2019–2022. We compute that 6323 people met this requirement during the survey period according to publication data in the ACL Anthology, meaning we have survey responses from about 5% of the total. For the rest of this paper, we restrict all reported results to this subset.

Full question text and results for all demographics are available in Appendix B.

3.1 Basic Demographics

Figure 1 shows location and gender statistics. Survey respondents are mostly men (67%) and mostly from the United States (58%). To get a sense of biases in our results, we compare to official statistics from the ACL.

Location The ACL publishes statistics about the countries of origin of its members. While ACL members are not the same population as our target demographic, we can use them as a rough

imum Impact Fund (https://www.givewell.org/maximum-impact-fund), $830 to GiveDirectly (https://www.givedirectly.org/), and $1,140 to the Distributed AI Research Institute (https://www.dair-institute.org/support). 23 respondents (5%) did not provide an answer to this question, so we did not make donations on their behalf.

proxy to assess for geographic bias. Comparing to the most recent available statistics[2] suggests that the United States is heavily overrepresented in our respondent population (58% > 35%), while Asia/Pacific is underrepresented (8% < 26%). Asian countries with large ACL contingents were particularly underrepresented, including China (3% < 9%), India (1% < 5%), and Japan (0.8% < 5%). We suspect this may largely be due to biases in our survey distribution methods (particularly Twitter and our personal networks).

Gender For gender, we compare to ACL-published diversity statistics, computed from the attendee population at ACL 2017 in Vancouver, Canada.[3] While our gender distribution is skewed, with 67% men and 25% women, it seems to roughly match the ACL population.

Underrepresented Minorities We ask respondents if they consider themselves to be part of an underrepresented minority group in NLP, to which 26% report *Yes*, 63% report *No*, and 11% *Prefer not to say*.

3.2 Career

Job Sector 73% of respondents are from academia, with 22% from industry and 4% in non-profit or government jobs. Asked if their job is "publication-oriented," 89% answer *Yes*, including 62% of respondents in industry, while 9% answer *No*.

Seniority Faculty and senior managers form a plurality of respondents at 41%, while 23% are junior professionals (including postdocs), 33% are PhD students, and 2% are Masters students or undergraduates. Breaking these down by sector, the largest group was academic PhD students (32%)[4], followed by faculty (29%), senior managers (10%), postdocs or other academic researchers (10%), and junior researchers in industry (10%).

Since ACL publishes statistics on the number of student memberships and regular memberships, we can also compare these numbers to the ACL population: 35% of respondents are students, compared to 53% of ACL members as of summer 2021. While students are underrepresented in our results compared to ACL members as a whole, it is not clear whether they are underrepresented compared to our target demographic of recently published authors.

We also ask respondents to report the year they published their first research work in NLP. 20% of respondents reported a year prior to 2009, with the earliest being 1985. Another 20% were in 2010–2014, 20% in 2015–2017, 20% in 2018–2019, and 10% in 2019–2022, with the rest declining to answer. See Appendix B, Figure 13 for more details.

Subfields We ask respondents to report all subfields that fit their work from the last 3 years, providing a list derived from the tracks at recent *CL conferences (EMNLP 2021, ACL 2022, and NAACL 2022). The most common responses are machine learning for NLP (39%), interpretability and analysis (31%), generation (28%), resources and evaluation (27%), and machine translation and multilinguality (26%). Coverage of the given subfields is broad, with each being marked by at least 20 respondents (6%). Full results are in Appendix B, Figure 16.

Research Activity Respondents are highly active at ACL events, with 96% saying they have attended an ACL event in the last 3 years. Asked about their total number of peer-reviewed publications related to NLP, 13% report 1–4, 47% report 5–20, and 39% report 21 or more.

3.3 Other Information

Twitter Use A large majority of respondents use Twitter, with 18% saying *I routinely post*, 58% saying *I follow but don't often post*, and 21% saying *I don't have a Twitter account, rarely look at it, or don't use it for research or NLP related content*. While we don't know what proportion

[2] https://www.aclweb.org/adminwiki/images/f/f4/Memberships_2021_by_Country_SUMMER.pdf

[3] https://www.aclweb.org/portal/content/acl-diversity-statistics

[4] The remaining 1% of PhD students reported being in industry or declined to report their job sector.

	1. 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 it.** Author anonymity during review is valuable enough to warrant restrictions on the dissemination of research that is under review.
	2. 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.
	3. 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.
	4. 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.
	5. Promising Research Programs
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. 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 2: All survey questions with agree/disagree answers. Only the full statements (not the bolded summaries) were shown to respondents. Besides these, the survey also asked for an opinion on likely sources of future advances in NLP (§4.7) as well as respondent demographics (§3, Appendix B).

of our target demographic uses Twitter, it seems likely that this is one of the largest sources of bias in our respondent population. 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.

Finding the Survey Asked how they heard about the survey, 37% reported hearing about it through Twitter, 21% from the email we sent through the ACL Member Portal, 19% from another mailing list, a Slack channel, or other messaging forum, 15% through word of mouth or personal communication, and 3% through an advertisement at ACL 2022, where we put up flyers and passed out stickers. Again, this suggests that our results are likely skewed towards Twitter users.

Meta-Question Confidence We ask respondents to report their confidence in their meta-question predictions. 5% report being *Completely confident*, 32% *Fairly confident*, 34% *Somewhat confident*, 13% *Slightly confident*, 13% *Not at all confident*, and 2% *Prefer not to say*. Between these groups, the *Completely confident* had the highest mean absolute error on their meta-question predictions (20.9%), though this was only significantly ($p < 0.05$) different from the *Slightly confident* group (18.5%), which performed best, despite abstaining from predictions in fewer cases (5%) than the completely confident group (11%). We considered removing meta-question answers from less confident respondents for the purpose of reporting results, but since there are few statistically significant differences and the less confident respondents actually performed better, we include all meta-question responses in the rest of this work for the sake of simplicity.

4 Results

For reference, all survey questions which took agree/disagree responses are shown in Figure 2. In the rest of this section, we will discuss the results of each section of the survey in detail.

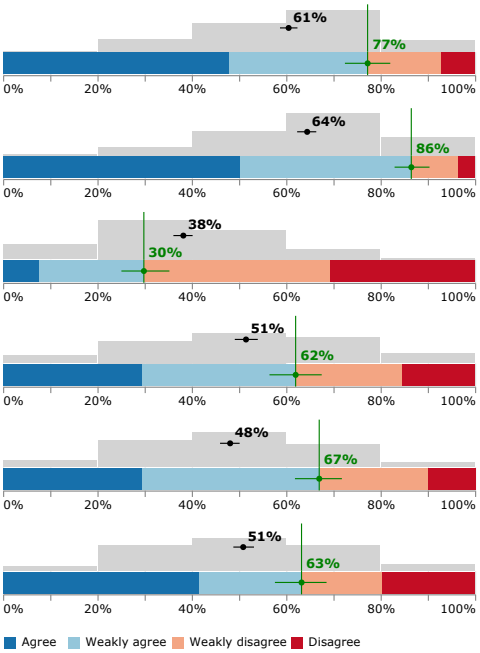
For each question (for example, in Figure 3), we display the proportion of Agree, Weakly agree, Weakly disagree, and Disagree answers in a band along the bottom of the visualization. These percentages exclude those who gave one of the other answers (Insufficiently informed on the issue, Question is ill-posed, or Prefer not to say), which were relatively rare (<20% of responses for all questions, <10% for 75% of questions; see Appendix A). The vertical green line shows the total percentage who Agree or Weakly agree with the statement, which was the value we asked survey respondents to predict in the meta-questions. The gray bars show the distribution of answers to the meta-questions, 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 comparison can be found online at `https://nlpsurvey.net/results/`.

4.1 State of the Field (Figure 3)

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

Industry's 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,



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.

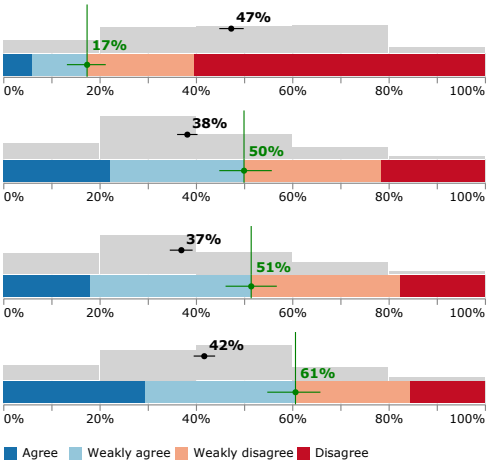
1-6. Author anonymity is worth it
Author anonymity during review is valuable enough to warrant restrictions on the dissemination of research that is under review.

Figure 3: *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.

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 pretraining (Devlin et al., 2019; Radford et al., 2019) draws away from other research agendas. Respondents underpredict 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 focusing on incorporating interdisciplinary insights as opposed to raw scaling, or problem formulation and task design—other under-predicted views, as we will see in §4.2, §4.5, and §4.7).

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, Eventually (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% AGREEing. 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).



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.

Figure 4: *Scale, Inductive Bias, and Adjacent Fields.*

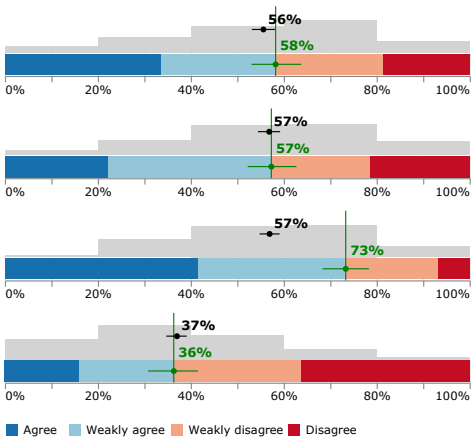
Dubious Science (Q1-5) A majority agrees that most NLP work is of "dubious scientific value" (67%). Respondents expressed uncertainty over what should count as "dubious," as well as concerns about who determines the value of research. On one hand, such research could refer to work which is fundamentally 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 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: 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.

4.2 Scale, Inductive Bias, and Adjacent Fields (Figure 4)

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 benefits from incorporating linguistic structure or expert-designed inductive biases to be superseded by learning mechanisms operating on fewer, more general principles if they have enough training data and model capacity. While the success of deep learning and large language models may be taken as supporting evidence for the Bitter Lesson, we find that the community has bought into the Lesson far less than it thinks it has.



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 5: *Artificial general intelligence (AGI) and major risks*.

Support for scaling maximalism is greatly overestimated (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 would be sufficient to practically solve any important problem in NLP (Q2-1). Overall, this is seen as a controversial issue, 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 cognitive 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.

4.3 AGI and Major Risks (Figure 5)

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 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 split near the middle, 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 some 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.

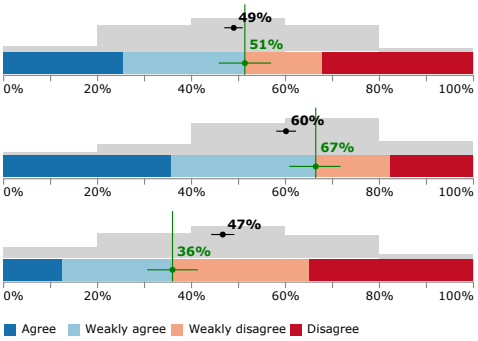
Revolutionary and catastrophic outcomes are a concern (Q3-3, Q3-4) 73% of respondents agree that labor automation from AI could plausibly lead to revolutionary societal change in this century, on at least the scale of the Industrial Revolution (Q3-3). This points to a common reason why those who agree 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 §5). But it's worth noting that a significant fraction of respondents (23%) agree with the prospect of revolutionary change (Q3-3) while disagreeing with the importance of AGI, suggesting that discussions about long-term or large-scale impacts of our work in NLP may not need to be tied up in the AGI debate.

About a third (36%) of respondents agree that it is plausible that AI could produce catastrophic outcomes in this century, on the level of all-out nuclear war (Q3-4). While this is a much smaller proportion than those who expect revolutionary societal change (Q3-3), the stakes are extremely high and a substantial minority expressing concern about such outcomes indicates that a deeper discussion of such risks may be warranted in the NLP community. While we do not ask about specific ways in which 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 present-day track record of disproportionate harms 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 also 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.

4.4 Language Understanding (Figure 6)

The question of whether language models understand language has been the subject of some debate in the community (Bender & 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 §4.3, 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



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 6: *Language Understanding*.

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 an estimate of 49% agreement. Many more people (67%) agree once the model has access to multimodal data (images, etc.). As with the importance of AGI (§4.3), 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 that some may believe that 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.

4.5 PROMISING RESEARCH PROGRAMS (FIGURE 7)

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 & 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 & Wallace, 2019; Wiegreffe & 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 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%.

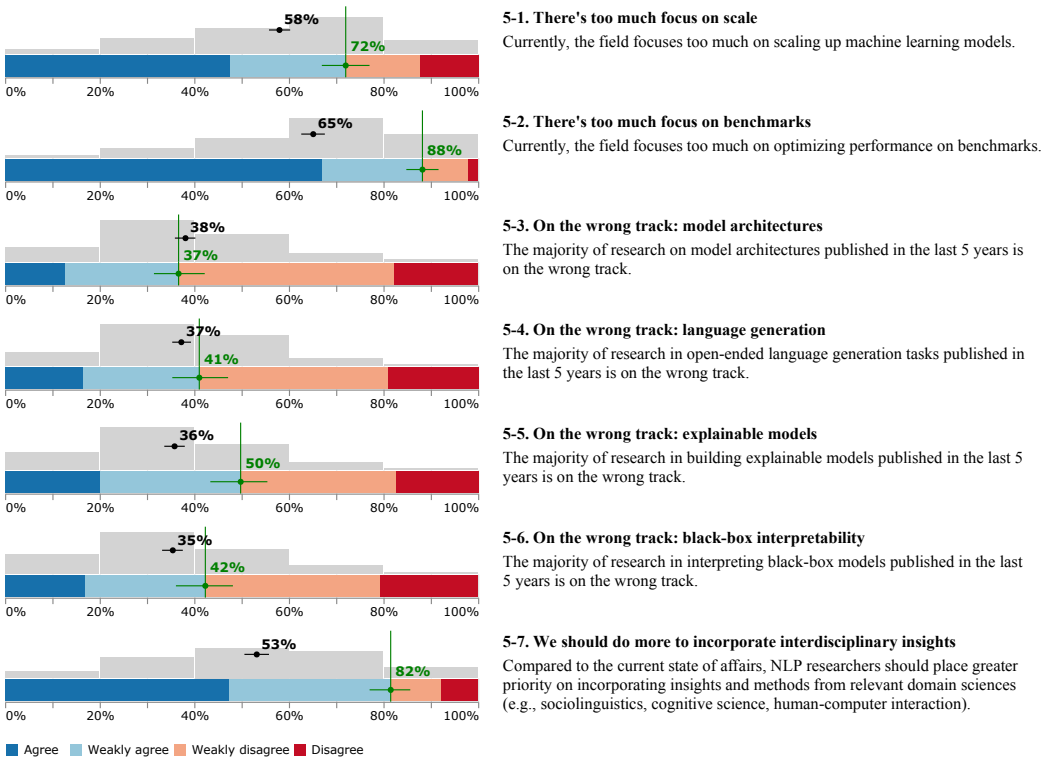


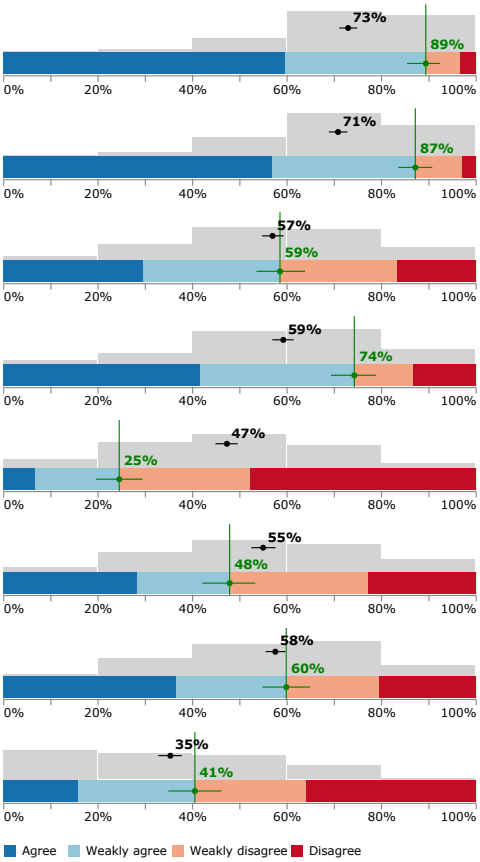
Figure 7: *Promising Research Programs*.

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.



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 8: *Ethics*.

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.

4.6 Ethics (Figure 8)

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 day (89%, Q6-1) and going into the future (87%, Q6-2). This strong endorsement of NLP's future impact stands in contrast with more substantial worries about catastrophic outcomes (36%, Q3-4). While the views are anticorrelated,[5] 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 will be high in the near future when it comes to the impact of NLP research.

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

[5] 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.

system can be used, how it is released, etc.). Our intent with the question is to get a sense of the 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, [6] which is adopted by the ACL [7] 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 agreement 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 overpredicted 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 machine 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 physiognomy (Agüera 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 this 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 different 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.

[6] https://www.acm.org/code-of-ethics
[7] https://www.aclweb.org/portal/content/acl-code-ethics

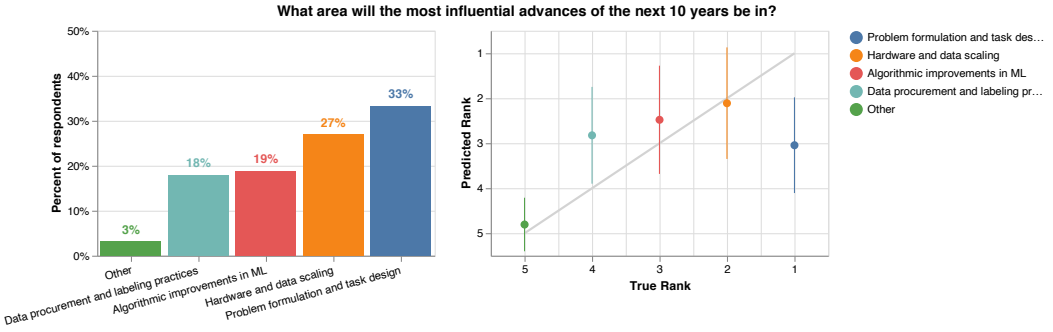


Figure 9: *Likely Sources of Future Advances*. The left shows the distribution of answers, and the right shows predicted ranks (mean and standard deviation) of each answer relative to its true rank.

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 respondents' 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 general skepticism of government regulation.

4.7 Likely Sources of Future Advances (Figure 9)

In addition to the agree/disagree questions that constitute most of the survey, we also 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).[8]

The results 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 (although noisy). 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

Having each survey respondent provide opinions on a wide range of issues gives us the opportunity to examine the relationships between these opinions: Which sets of beliefs often come together, and which don't? Also, are there demographic characteristics that correlate with certain beliefs? 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).

[8] 20% of respondents rank "Other" on top (rank 1) for this question, even though very few actually provided it as their answer. It seems to us that these respondents probably 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 D, Figure 18).

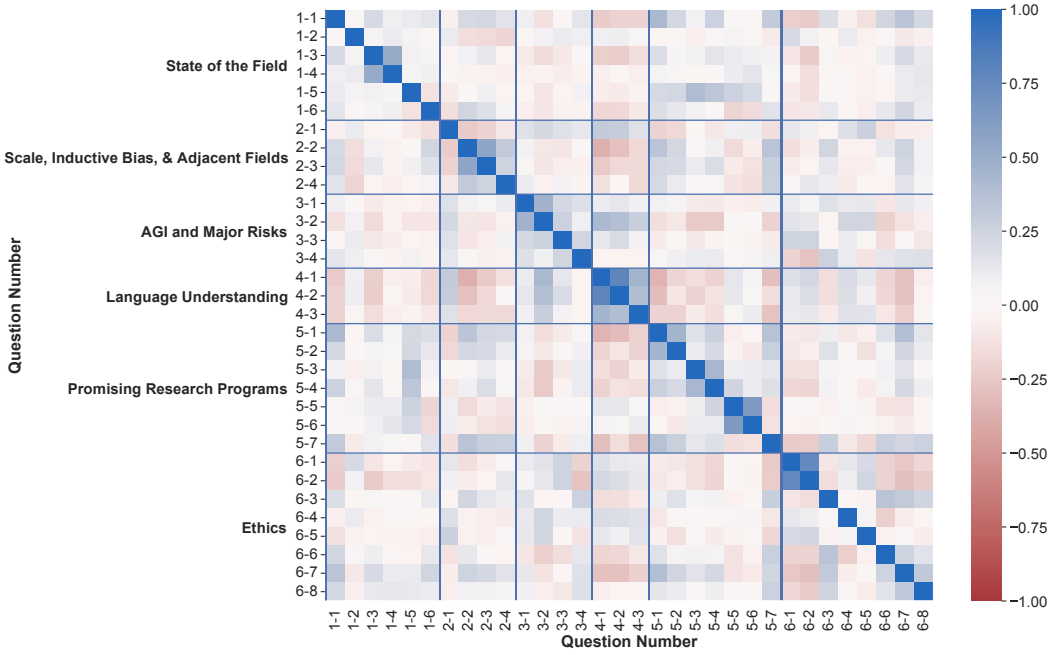


Figure 10: Spearman correlations between answers to all pairs of agree/disagree questions. Lines separate sections of the survey. Question numbers are given in Figure 2.

5.1 Question Correlations

Spearman (rank-order) correlations between agree/disagree questions are shown in Figure 10. To compute these, we order the response values as [Disagree, Weakly disagree, Other, Weakly agree, Agree], where Other includes Insufficiently informed on the issue, Question is ill-posed, and Prefer not to say.[9] Unsurprisingly, correlations tend to be stronger within a single section of the survey — for example, perspectives on linguistic structure and inductive bias (§4.2), AGI (§4.3), language understanding (§4.4), and NLP's net impact on society (§4.6, Q6-1, Q6-2). Beyond these, Table 1 shows the highest-magnitude correlations between questions in different survey 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. This 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, $\rho_s = 0.30$), 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, $\rho_s = -0.30$). 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

[9] Even though the Other answers may not be meaningfully "in between" agreeing and disagreeing, we find that running the analysis excluding Other options produces nearly identical results, so we keep them in order to apply the analysis to all of the data.

Q1	Q2	ρ_s
Q**1-1.** Private firms have too much influence.	Q**5-1.** There's too much focus on scale.	**+0.43**
Q**3-2.** Recent progress is moving us towards AGI.	Q**4-1.** LMs understand language.	**+0.42**
Q**1-5.** Most of NLP is dubious science.	Q**5-3.** On the wrong track: model architectures.	**+0.40**
Q**5-1.** There's too much focus on scale.	Q**6-7.** Carbon footprint is a major concern.	**+0.38**
Q**1-5.** Most of NLP is dubious science.	Q**4-2.** Multimodal models understand language.	**+0.38**
Q**2-2.** Linguistic structure is necessary.	Q**4-1.** LMs understand language.	**-0.38**
Q**4-1.** LMs understand language.	Q**5-1.** There's too much focus on scale.	**-0.35**
Q**1-1.** Private firms have too much influence.	Q**6-7.** Carbon footprint is a major concern.	**+0.35**
Q**2-2.** Linguistic structure is necessary.	Q**5-7.** We should incorporate more interdisciplinary insights.	**+0.35**
Q**2-2.** Linguistic structure is necessary.	Q**5-1** There's too much focus on scale.	**+0.34**
Q**1-5.** Most of NLP is dubious science.	Q**5-4.** On the wrong track: language generation.	**+0.33**
Q**1-1.** Private firms have too much influence.	Q**5-7.** We should incorporate more interdisciplinary insights.	**+0.30**
Q**4-2.** Multimodal models understand language.	Q**5-1.** There's too much focus on scale.	**-0.30**
Q**4-1.** LMs understand language.	Q**5-7.** We should incorporate more interdisciplinary insights.	**-0.30**
Q**2-1.** Scaling solves practically any important problem.	Q**4-1.** LMs understand language.	**+0.30**
Q**2-2.** Linguistic structure is necessary.	Q**4-2.** Multimodal models understand language.	**-0.30**
Q**2-1.** Scaling solves practically any important problem.	Q**4-2.** Multimodal models understand language.	**+0.29**
Q**4-2.** Multimodal models understand language.	Q**6-7.** Carbon footprint is a major concern.	**-0.28**
Q**3-2.** Recent progress is moving us towards AGI.	Q**4-3.** Text-only evaluation can measure language understanding.	**+0.28**
Q**5-7.** We should incorporate more interdisciplinary insights.	Q**6-3.** It is unethical to build easily misusable systems.	**+0.28**

Table 1: Top 20 Spearman correlations (ρ_s) between pairs of questions from distinct categories (i.e., with distinct first numbers). Correlations $|\rho_s| > 0.11$ are significant with $p < 0.05$.

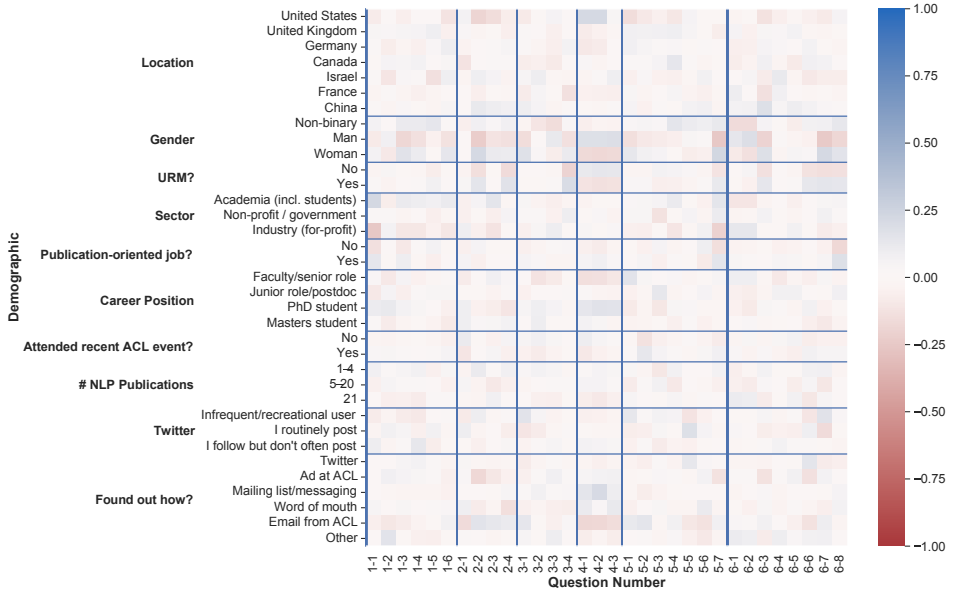


Figure 11: Spearman correlations between membership demographic groups and answers to agree/disagree questions. Lines separate survey sections and demographic variables. "URM" stands for under-represented minority. We only show demographic values with> 5 respondents. Question numbers are given in Figure 2.

and methods from linguistics, cognitive science, or other domain sciences. It is worth emphasizing, though, that none of our measured correlations here are extremely strong; 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 Figure 11, 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.

Demographic	Question	ρ_s
Sector: Industry (for-profit)	Q**1-1.** Private firms have too much influence.	**-0.25**
Gender: Woman	Q**5-7.** We should incorporate more interdisciplinary insights.	**+0.24**
Location: United States	Q**4-2.** Multimodal models understand language.	**+0.22**
Location: United States	Q**4-1.** LMs understand language.	**+0.22**
Sector: Academia (including students)	Q**1-1.** Private firms have too much influence.	**+0.21**
Gender: Man	Q**6-2.** NLP's future net impact will be good.	**+0.21**
Sector: Industry (for-profit)	Q**5-7.** We should incorporate more interdisciplinary insights.	**-0.21**
Gender: Woman	Q**6-7.** Carbon footprint is a major concern.	**+0.20**
Gender: Woman	Q**2-2.** Linguistic structure is necessary.	**+0.19**
Under-represented Minority: Yes	Q**3-4.** AI decisions could cause nuclear-level catastrophe.	**+0.19**

Table 2: Top 10 Spearman correlations (ρ_s) between membership in a demographic 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%)	
Q**4-1.** LMs **do** understand language	+0.36	Q**3-1.** AGI **is** important	+0.48
Q**4-2.** Multimodal models **do** understand	+0.29	Q**3-2.** We **are** stepping to AGI	+0.38
Q**6-7.** Carbon **isn't** a major concern	-0.29	Q**6-3.** Easy misuse **is** unethical	+0.28
Q**5-1.** There **isn't** too much focus on scale	-0.28	Q**6-7.** Carbon **is** a major concern	+0.26
Q**2-2.** Linguistic structure **isn't** necessary	-0.26	Q**3-3.** Revolutionary change **is** plausible	+0.24
Deep Learning Pessimism (5.2%)		**Jaded Empiricism (4.0%)**	
Q**5-5.** **Wrong** track (explainability)	+0.33	Q**6-6.** **Not** unethical to predict psych.	-0.37
Q**5-6.** **Wrong** track (interpretability)	+0.32	Q**1-5.** Most NLP **is** dubious science	+0.30
Q**1-5.** Most NLP **is** dubious science	+0.28	Q**5-5.** **Wrong** track (explainability)	+0.27
Q**3-4.** Catastrophic risk **is** plausible	+0.25	Q**5-6.** **Wrong** track (interpretability)	+0.24
Q**6-2.** NLP **isn't** good (future)	-0.23	**Has** 21+ Pubs.	+0.24

Table 3: The top four components from running PCA on survey responses and demographic data, with human-written 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.

Membership in demographic groups does not strongly correlate with answers to any questions in the survey. The strongest correlation is $\rho_s = -0.25$ (Table 2), smaller than the top 20 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 answer many questions differently**, especially regarding ethics and belief in the value of interdisciplinary research. In addition, **under-represented minorities are more likely to agree that AI could have catastrophic consequences this century**. Perhaps this reflects a worry that more powerful AI systems would have especially negative impacts on under-represented minorities, or that minorities would be negatively affected before society as a whole is significantly affected.

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 $[-1, 1]$: {Disagree $\to -1$, Weakly disagree $\to -0.5$, Other $\to 0$, Weakly agree $\to 0.5$, Agree $\to 1$}, where Other includes Question is ill-posed, Insufficiently informed on the issue, and Prefer not to say. For demographics, we treat every answer choice as a 0/1 binary variable as in §5.2. 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"), explaining 11.7% of the variance in the data, and 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.

As found in §5.2, demographic features are not as explanatory of the data as the agree/disagree questions are. Accordingly, they are not among the questions most strongly associated with the top clusters. The exception is the fourth component ("Jaded Empiricism"), for which one of the strongest associations is having a large number of publications, though this component only explains 4.0% of the total variance.

6 Related Work

This work is directly inspired by the 2009 PhilPapers Surveys (Bourget & Chalmers, 2014), a survey and metasurvey of professional philosophers designed to discover and compare both their philosophical beliefs and their sociological beliefs about the philosophy community. More generally, introspective surveys to get a sense of the demographics and opinions of a field are common in many disciplines, such as economics (Fuchs et al., 1998; Illge & Schwarze, 2009; Frey et al., 2010, *inter alia*) and physics (Sivasundaram & Nielsen, 2016). We design our survey to fill a similar role for NLP, with a special focus on controversies where the field's perception of its views may not match reality, due for example to the influence of social media or a small number of dominant voices.

The NLP Community Metasurvey overlaps slightly with other surveys of the NLP community, which are typically organized by ACL committees and focused on specific, timely issues related to the organization policy or logistics. Issues covered by such surveys include preprint posting practices (Foster et al., 2017), experiences with rolling review infrastructure (Schütze, 2022), ethical stances (Fort & Couillault, 2016; Benotti et al., 2022), attitudes towards energy use and environmental impacts (Arase et al., 2021), and language diversity in NLP research (Diab, 2022). Instead of focusing on a specific issue, our aim with the NLP Community Survey is to provide a broad sense of the distribution of views (and meta-views) held by NLP researchers. To do this, we cover a range of topics with existing debate in the literature (see the subsections of §4 for references). This is related to the the NLP Scholar project, which analyzes broad demographic (Mohammad, 2020c; 2019), publication (Mohammad, 2020b), and citation (Mohammad, 2020a) trends in NLP and their interplay.

7 Discussion

With the NLP Community Metasurvey, we have made concrete numbers of 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 ($\sim$50%) under uncertainty.

There are other biases to keep in mind when interpreting our results: people in the United States are overrepresented in our respondent population relative to ACL members as a whole, and senior researchers and academics are probably overrepresented as well (Appendix B), not to mention unmeasured population biases based in the personal networks of the authors. Many of the 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 reasonable 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 and colleagues who provided feedback on the survey.

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.

REFERENCES

Blaise Agüera y Arcas, Margaret Mitchell, and Alexander Todorov. Physiognomy's new clothes, May 2017. URL https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety, 2016. URL https://arxiv.org/abs/1606.06565.

Yuki Arase, Phil Blunsom, Mona Diab, Jesse Dodge, Iryna Gurevych, Percy Liang, Colin Raffel, Andreas Rücklé, Roy Schwartz, Noah A. Smith, Emma Strubell, and Yue Zhang. Efficient NLP survey, 2021. URL https://www.aclweb.org/portal/content/efficient-nlp-survey.

Emily M. Bender and Alexander Koller. 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*, pp. 5185–5198, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463. URL https://aclanthology.org/2020.acl-main.463.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 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, pp. 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Lucia Benotti, Mark Drezde, Karën Fort, Pascale Fung, Dirk Hovy, Min-Yen Kan, Jin Dong Kim, Malvina Nissim, and Yulia Tsvetkov. 2022 ACL ethics survey, 2022.

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. Khattab, 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, Avanika 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. On the opportunities and risks of foundation models. *ArXiv*, 2021. URL https://crfm.stanford.edu/assets/report.pdf.

Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Inc., USA, 1st edition, 2014. ISBN 0199678111.

David Bourget and David J. Chalmers. On the conception and design of the PhilPapers Survey. URL https://philpapers.org/surveys/designthoughts.html. Retrieved August 2022.

David Bourget and David J. Chalmers. What do philosophers believe? *Philosophical Studies*, 170 (3):465–500, 2014. doi: 10.1007/s11098-013-0259-7.

David Bourget and David J. Chalmers. Philosophers on philosophy: The PhilPapers 2020 survey, 2020.

Samuel R. Bowman and George Dahl. 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*, pp. 4843–4855, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.385. URL https://aclanthology.org/2021.naacl-main.385.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

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 Sankaranarayana Pillai, 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. PaLM: Scaling language modeling with pathways. *CoRR*, 2022. doi: 10.48550/ARXIV.2204.02311. URL https://arxiv.org/abs/2204.02311.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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), pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Mona Diab. Language diversity in our ACL community, 2022. URL https://forms.office.com/pages/responsepage.aspx?id=DQSIkWdsW0yxEjajBLZtrQAAAAAAAAAAAAN__oJ4BAFUNlVMMFpOVlkzM1FSRlZaUlo0VEI4OEI5Ty4u.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3719–3728, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1407. URL https://aclanthology.org/D18-1407.

Karën Fort and Alain Couillault. Yes, we care! Results of the ethics and natural language processing surveys. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 1593–1600, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1252.

Jennifer Foster, Marti Hearst, Joakim Nivre, and Shiqi Zhao. Report on ACL survey on preprint publishing and reviewing. Association for Computational Linguistics, 2017.

Bruno S Frey, Silke Humbert, and Friedrich Schneider. What is economics? Attitudes and views of German economists. *Journal of Economic Methodology*, 17(3):317–332, 2010.

Victor R Fuchs, Alan B Krueger, and James M Poterba. Economists' views about parameters, values, and policies: Survey results in labor and public economics. *Journal of Economic Literature*, 36 (3):1387–1425, 1998.

Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems, 2019. URLhttps://arxiv.org/abs/1906.01820.

Lydia Illge and Reimund Schwarze. A matter of opinion—How ecological and neoclassical environmental economists and think about sustainability and economics. *Ecological Economics*, 68 (3):594–604, 2009.

John P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2, 2005.

Sarthak Jain and Byron C. Wallace. 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)*, pp. 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357.

Andrey Kurenkov. Lessons from the PULSE model and discussion. *The Gradient*, 2020.

William Merrill, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. Provable Limitations of Acquiring Meaning from Ungrounded Form: What Will Future Language Models Understand? *Transactions of the Association for Computational Linguistics*, 9:1047–1060, 09 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00412. URL https://doi.org/10.1162/tacl_a_00412.

Saif M. Mohammad. The state of NLP literature: A diachronic analysis of the ACL Anthology. *arXiv preprint 1911.03562*, 2019.

Saif M. Mohammad. Examining citations of natural language processing literature. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5199–5209, Online, July 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.464. URL https://aclanthology.org/2020.acl-main.464.

Saif M. Mohammad. NLP scholar: A dataset for examining the state of NLP research. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 868–877, Marseille, France, May 2020b. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.109.

Saif M. Mohammad. Gender gap in natural language processing research: Disparities in authorship and citations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7860–7870, Online, July 2020c. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.702. URL https://aclanthology.org/2020.acl-main.702.

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. Do transformer modifications transfer across implementations and applications? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5758–5773, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.465. URL https://aclanthology.org/2021.emnlp-main.465.

David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David R. So, Maud Texier, and Jeff Dean. The carbon footprint of machine learning training will plateau, then shrink. *Computer*, 55(7):18–28, 2022. doi: 10.1109/MC.2022.3148714.

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

Inioluwa Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 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)*, 2021.URL https://openreview.net/forum?id=j6NxpQbREA1.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green AI. *Communications of the ACM*, 63(12):54–63, 2020. doi: 10.1145/3381831. URL https://doi.org/10.1145/3381831.

Hinrich Schütze. Survey on the future of reviewing, 2022. URL https://www.cis.lmu.de/~hs/acl22/panel/survey.html.

Sujeevan Sivasundaram and Kristian Hvidtfelt Nielsen. Surveying the attitudes of physicists concerning foundational issues of quantum mechanics. *arXiv preprint 1612.00676*, 2016.

Rich Sutton. The bitter lesson, 2019. URL http://www.incompleteideas.net/IncIdeas/BitterLesson.html.

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

Sarah Wiegreffe and Yuval Pinter. 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)*, pp. 11–20, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1002. URL https://aclanthology.org/D19-1002.

A Overview of Responses

A full overview of the responses to agree/disagree questions is shown in Figure 12. The Other answers were fairly uncommon—never above 20% of all answers—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}).

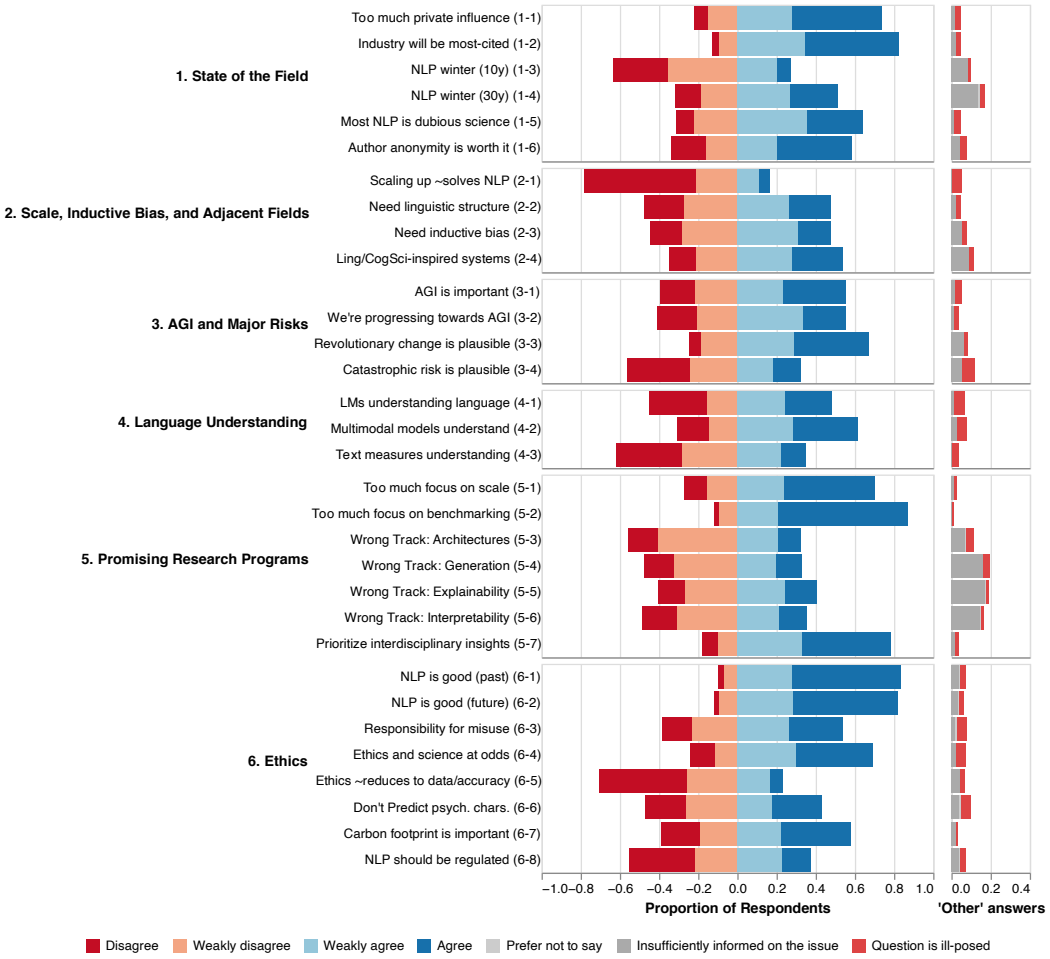


Figure 12: Full overview of responses.

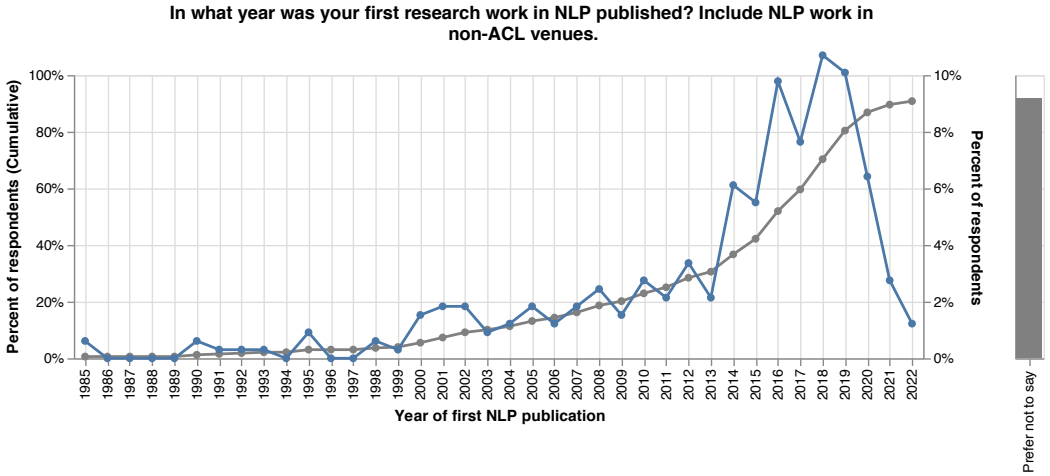


Figure 13: Respondents' year of first NLP publication.

B Detailed Demographics

Figures 14–17 show full results for the demographics questions. Results are restricted to those in the target demographic, i.e., with at least 2 *CL publications in the last three years. Numeric labels for percentages below 5% are omitted from the charts for space.

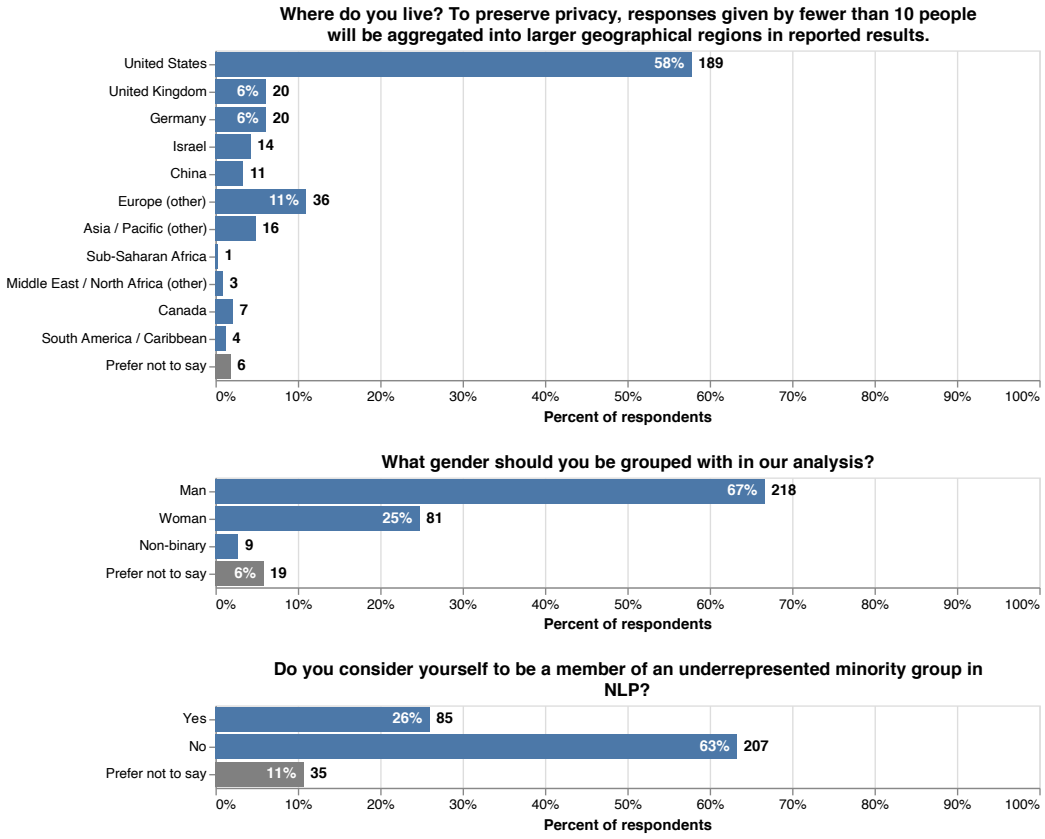


Figure 14: Basic demographics.

C 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.

D Data Cleaning and Postprocessing

Likely Sources of Future Advances As mentioned in §4.7, 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

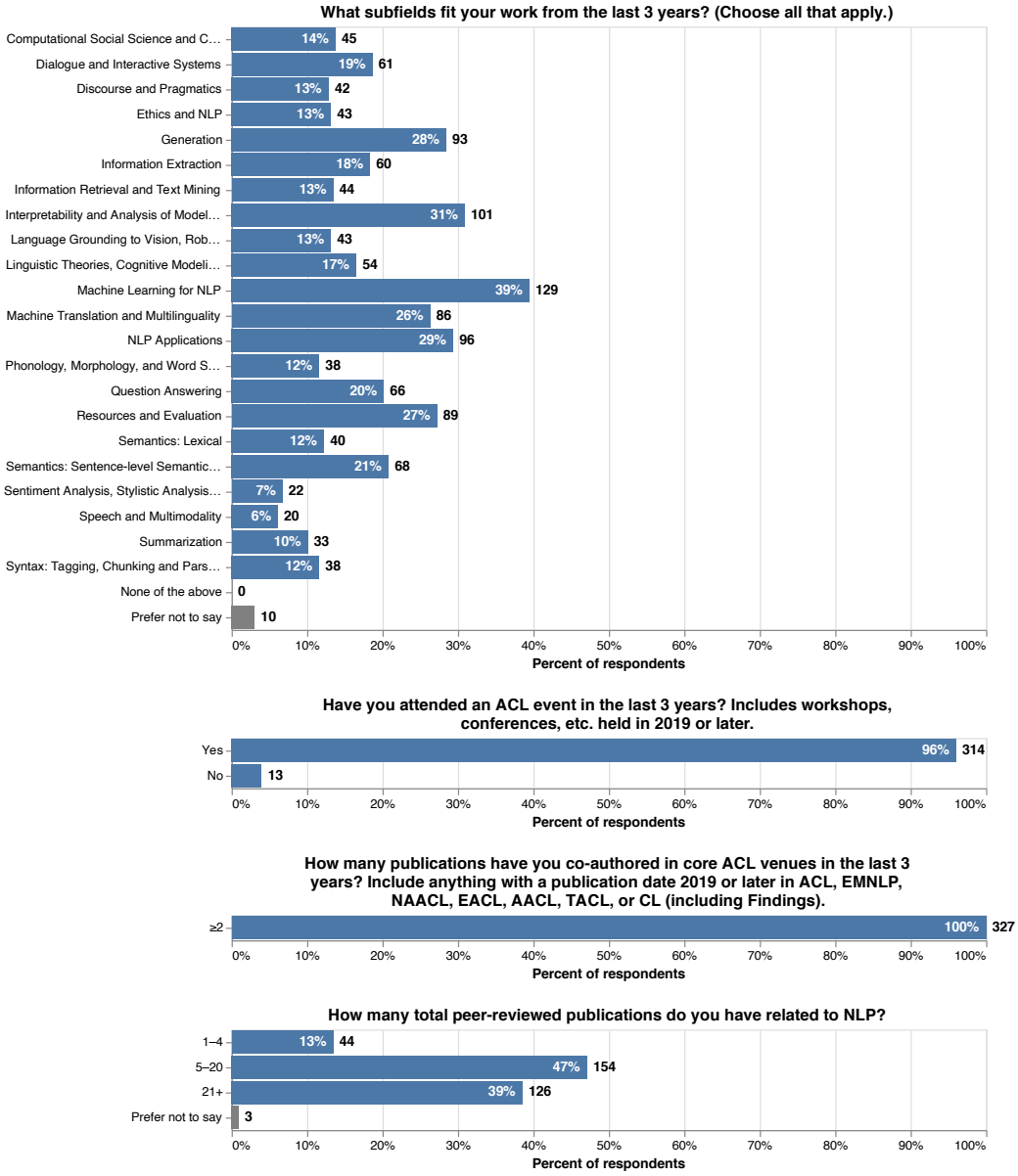


Figure 15: Respondents' research activities.

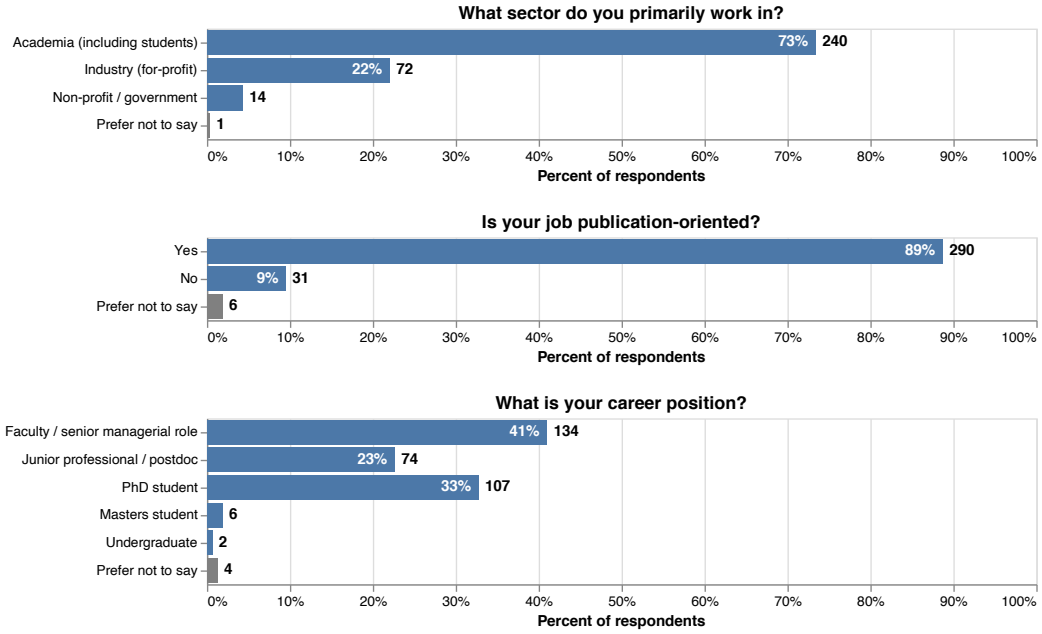


Figure 16: Career demographics.

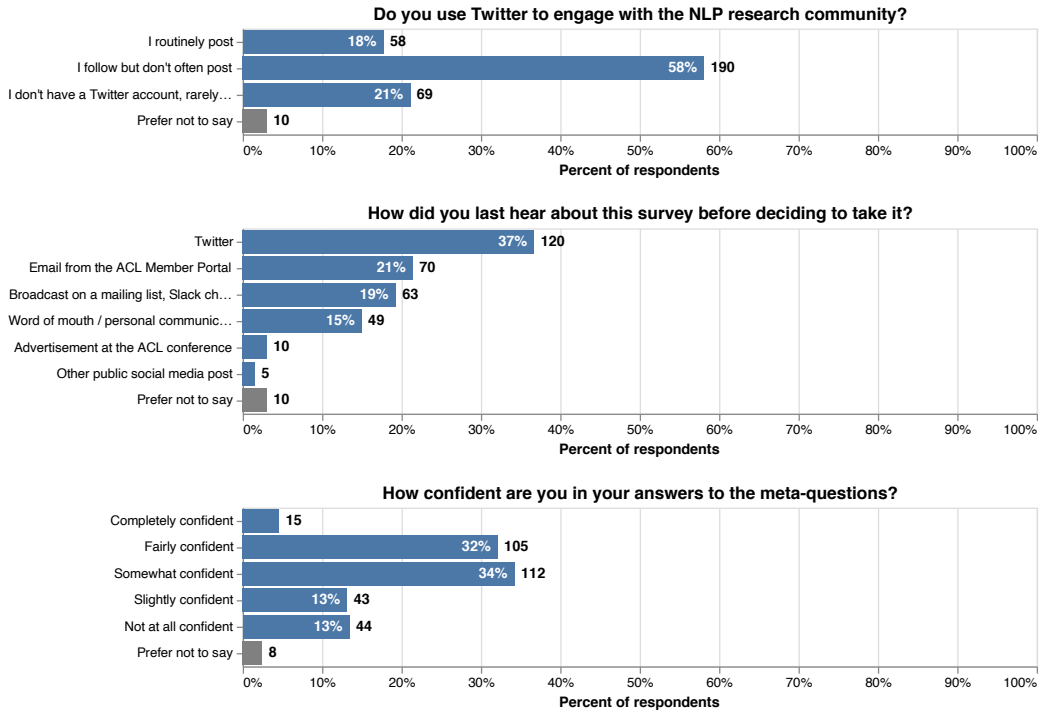


Figure 17: Other information provided by respondents.

of analysis. Besides changing the "Other" statistics, this does not seem to have a noticeable effect on overall trends (Figure 18).

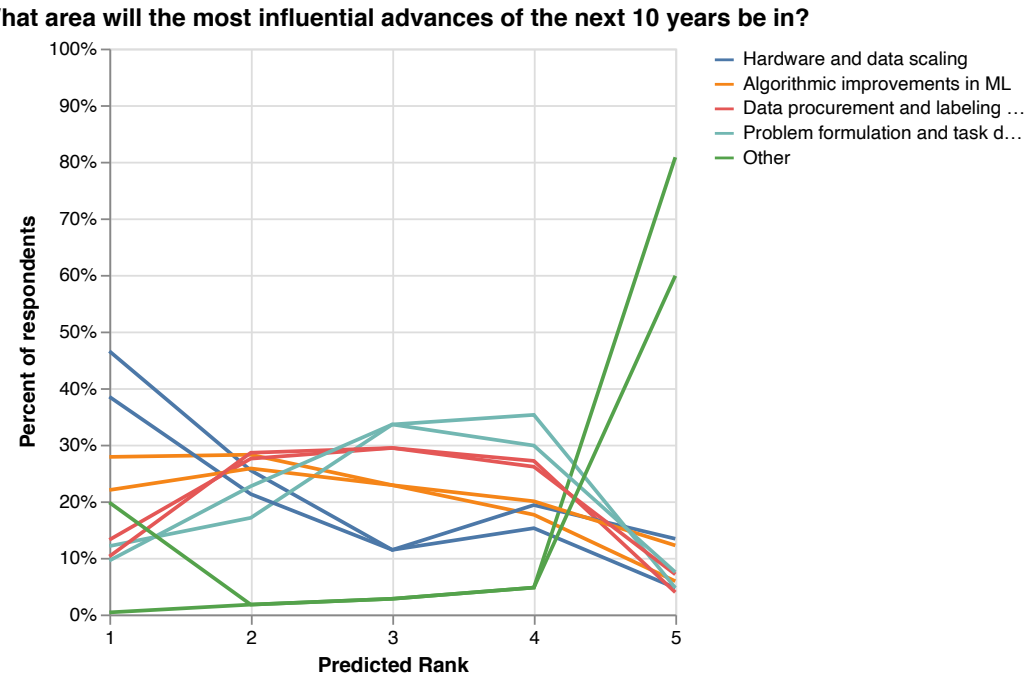


Figure 18: We postprocessed the predicted rankings for likely sources of future advances by reversing the rankings given by all respondents who placed "Other" first (20% of respondents). The rankings for the unadjusted data are shown in the faded lines; besides the change to the "Other" line, overall trends are the same.

D.1 Survey Instructions

The survey instructions are reproduced in full below, and examples of how it looks in the browser are given in Figure 19.

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. This was inspired by the PhilPapers surveys (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) 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.

You can reach us with questions and concerns at nlp-metasurvey-admin@nyu.edu.

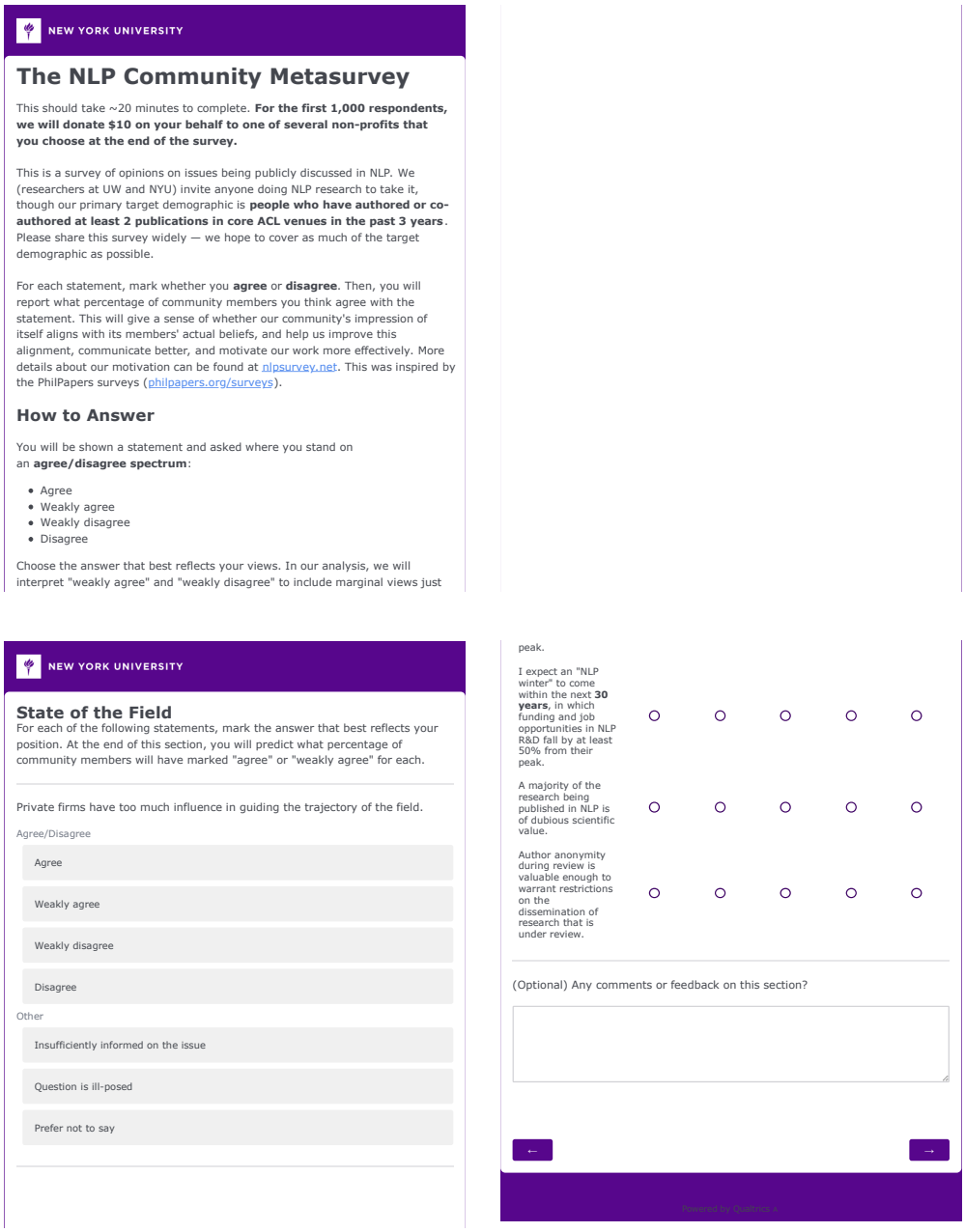


Figure 19: How the survey looks to respondents in a web browser.