Research Community Perspectives on "Intelligence" and Large Language Models

Anonymous ACL submission

Abstract

Despite the widespread use of "artificial intelligence" (AI) framing in Natural Language Processing (NLP) research, it is not clear what researchers mean by "intelligence". To that end, we present the results of a survey on the notion of "intelligence" among researchers and its role in the research agenda. The survey elicited complete responses from 303 researchers from a variety of fields including NLP, Machine Learning, Cognitive Science, Linguistics, Neuroscience. We identify 3 criteria of intelligence that the community agrees on the most: *generalization, adaptability, & reasoning*. Our results suggests that the perception of the current NLP systems as "intelligent" is a minority position (29%). Furthermore, only 16.2% of the respondents see developing intelligent systems as a research goal, and these respondents are more likely to consider the current systems intelligent.

1 Introduction

Whether machines can be considered intelligent has long been a question of interest to both researchers and industry. Researchers have a rich tradition of devising intelligence tests for machines, starting from the Turing test (Turing, 1950), and the industry increasingly markets "smart" and "intelligent" systems. A recent iteration of that is LLM-based systems, also referred to as "generative AI",[1] which have been suggested to show "sparks of artificial general intelligence" (Bubeck et al., 2023).

Irrespective of one's position on whether systems based on current machine learning technology can be intelligent, there is undeniably an increase in the phenomenon we refer to as *seeming intelligence*: the user's perception that a system is intelligent, based on their interaction with that system. This perception arises from the Eliza effect: humans have a strong tendency to attribute to their conversational partners, even simple chatbots, "all sorts of background knowledge, insights and reasoning ability" (Weizenbaum, 1966). Kim and Im (2023) conclude that the perception of intelligence in interactive systems is a key driver of anthropomorphization, irrespective of whether the system is presented via an avatar or not.

The tendency to anthropomorphize AI systems is not limited to the lay public or marketing by tech companies:[2] Mitchell (2021) points out that even

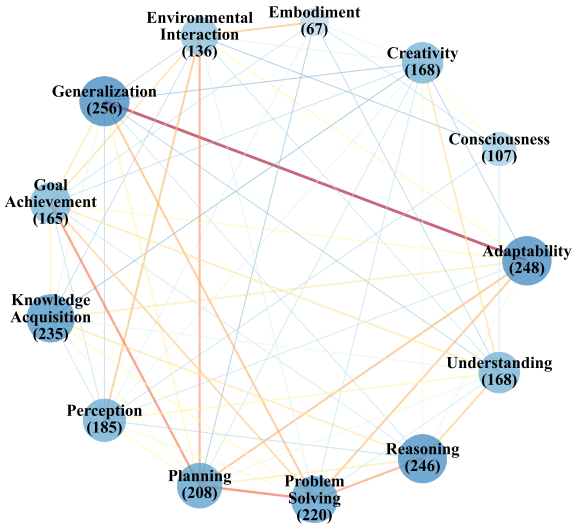


Figure 1: Correlation between criteria that the survey respondents selected as relevant for their notion of "intelligence". Darker edges indicate stronger correlations, larger nodes indicate higher relevance. Only edges with $\phi > |0.1|$ are shown.

[1]The term "Generative AI" is used for models of various modalities, but this study is concerned only with systems based on Large Language Models (LLMs), i.e., models trained on large text corpora that are used for transfer learning (Rogers and Luccioni, 2024). We use the term "AI" in line with its common use in the field, but we note that it entails that the term "intelligence" is applicable, which, as our results suggest, is not accepted by the majority of the respondents.

[2]To mention a few examples, OpenAI describe their base class of models as "*a set of models without instruction following that can* ***understand*** *(...) natural language and code*", and

researchers mislead themselves by using terms like "understanding", which have long been criticized as "wishful mnemonics" (McDermott, 1976). In particular, we have a longstanding tradition of benchmarks of "understanding" (e.g. Wang et al., 2018, 2020), "reasoning" (e.g. Vilares and Gómez-Rodríguez, 2019; Srivastava et al., 2023), "behaviors" (e.g. Chang and Bergen, 2023) and "capabilities" (e.g. Chen et al., 2024; Srivastava et al., 2023).

Given the above tendency towards anthropomorphization and the long history of "wishful mnemonics" in the academic literature, the goal of this study is to provide a snapshot of how researchers think about "intelligence" two years after the release of ChatGPT (OpenAI, 2022). We pose the following research questions:

RQ1 *What are key criteria for a system to be considered "intelligent", and to what degree do researchers in different fields, occupations, and career stages agree?*

RQ2 *What is the perception of the "intelligence" of current LLM-based systems such as ChatGPT?*

RQ3 *What role, if any, does the research agenda of respondents play in the notion of "intelligence"?*

Our results indicate a high level of coherence across groups in terms of the criteria that are relevant for intelligence. Independent of one's field of research, occupation, and career stage, most researchers believe that *generalization*, *adaptability*, and *reasoning* are key aspects of intelligence (**RQ1**, illustrated in Figure 1). We additionally observe a general skepticism regarding the attribution of intelligence to LLM-based systems, although it decreases for the future systems based on a similar technology (**RQ2**). For **RQ3**, we find that only a small subset of respondents consider it their goal to build intelligent technology, with the majority rather aiming to add to the scientific knowledge. Interestingly, the survey shows that those who consider 'creating intelligent systems' as their research goal are also more likely to attribute intelligence to the current systems (**RQ2**, **RQ3**).

GPT-4 is better "... *thanks to its broader general* ***knowledge*** *and advanced* ***reasoning*** *capabilities."* and finally the o1 class of models "***think*** *before they answer*" (OpenAI). Anthropic stated that the fundamental limitation of "**logical reasoning**" has already been solved with current systems (Anthropic).

2 Background

2.1 Defining "Intelligence"

According to Legg and Hutter, a fundamental problem in the field of AI is the fact that no one knows what "intelligence" is. Legg and Hutter (2007) provide a list of 72 definitions of intelligence from various fields, highlighting two key features as fundamental. First, intelligence is often ascribed to agents in the context of an environment. Second, many definitions emphasize the ability to adapt to novel situations.

As emphasized by their list, there are many different definitions of "intelligence" coming from a variety of research fields, which focus on different aspects. In the machine learning (ML) community, definitions tend to focus on skill acquisition (e.g. Chollet, 2019; LeCun, 2022). In contrast, psychological and dictionary definitions typically emphasize behavioral aspects and social dimensions (Legg and Hutter, 2007; Albus, 1991). The definitions, and tests, developed for *human* intelligence may be directly applied to machines, e.g. via benchmarks that purport to test mathematical reasoning, which is then taken as a proxy for intelligence. Gignac and Szodorai (2024) highlight this methodological tension, arguing for clearer nomenclature and AI-specific evaluation metrics.

2.2 Relevant NLP Research Directions

There are countless research directions relevant to the discussion of "intelligence" which span many fields, and which we cannot do justice in scope of this work. The remainder of this section provides pointers to three prominent directions within the field of NLP.

The meaning debate. There is an active debate on whether terms like "meaning" and "understanding" apply to LLMs, with prominent opponents (Bender and Koller, 2020; Merrill et al., 2021) and proponents (Li et al., 2024).

"Science of datasets". Benchmarks such as SuperGLUE (Wang et al., 2020), MMLU (Hendrycks et al.), and GSM8K (Cobbe et al., 2021) are used widely both for academic evaluation of LLM-based systems, and in their marketing. But how much we can deduce from the current benchmark results is up for debate. Ideally, the creation of benchmarks and training resources should consider the construct validity of the phenomena being modeled, and hence rely on specific definitions of their

target phenomena and representativeness of their samples (Schlangen, 2021; Raji et al., 2021), but we do not yet have solid methodology for doing this well for complex and underdefined constructs like "reasoning". Related research directions include documenting (Bender and Friedman, 2018; Gebru et al., 2020) and 'measuring' the existing datasets (Dodge et al., 2021; Mitchell et al., 2023), in particular to identify shortcuts in them (Gururangan et al., 2018; Gardner et al., 2021) and measure test contamination (Deng et al., 2024; Dong et al., 2024; Jacovi et al., 2023). Another relevant direction is creating adversarial datasets to 'catch' the models relying on shallow heuristics (e.g. McCoy et al., 2019).

Interpretability research. The field of interpretability (Belinkov and Glass, 2019; Rogers et al., 2020), including the mechanistic camp (Saphra and Wiegreffe, 2024; Olsson et al., 2022; Nanda et al., 2023), aims to identify the internal mechanisms of the models through which they arrive at their predictions. Assuming that we know what the "correct" reasoning steps should be, this could allow us to check whether the model adheres to them, or relies on shallow heuristics (e.g. Ray Choudhury et al., 2022).

3 Survey Structure and Distribution

The survey was developed using the SurveyXact platform[3]. We initially provide the introductory information regarding anonymity and informed consent, followed by questions related to basic demographic information relevant for the analysis. This is followed by questions pertaining to the respondents' perspectives on the notion of "intelligence". Responses were collected between November 2024 and January 2025. The survey form and response data (with coarse-grained geographic data for increased anonymity) is provided[4] in the supplementary materials of the submission, and a full list of questions can also be found in Appendix A.

The survey was distributed through social media (Bluesky and LinkedIn) as well as the mailing lists, the full list is available in the Appendix (Table 6). The mailing lists were selected to collect diverse responses from researchers in multiple fields in which the term "intelligence" is used, including NLP, ML, Cognitive Science, Philosophy, and Psychology. We additionally distributed the survey through internal mailing lists in research groups within the authors' networks (mostly NLP, Computational Social Science and Human-Computer Interaction).

[3] https://www.survey-xact.dk/

[4] Upon the publication of the paper all the accompanying data will be made available for research purposes under CC BY-NC 4.0 license.

Criterion	Sources
Prob. solving	Turing (1950); Mitchell (2021)
Knowledge acq.	Piaget (1947); Chollet (2019)
Adaptability	Legg and Hutter (2007); Pfeifer (2006)
Generalization	Marcus (2003); Tenenbaum et al. (2011)
Goal achievement	Piaget (1952); Van Gerven (2017)
Reasoning	Legg and Hutter (2007); Lake et al. (2017)
Understanding	Gardner (1983); Tenenbaum et al. (2011)
Perception	Smith and Gasser (2005); Van Gerven (2017)
Planning	Albus (1991)
Creativity	Wiggins (2020)
Env. interaction	Piaget (1952); Pfeifer and Scheier (2001)
Consciousness	Wiggins (2020)
Embodiment	Smith and Gasser (2005); Pfeifer (2006)

Table 1: The criteria of "intelligence" presented to the survey participants, and their sources (methodology described in subsection 4.2).

4 Question Design

This section provides a brief overview of the questions included in the survey and the rationale for including them. An overview of the questions is provided in Appendix A.

4.1 Demographics

The survey includes the following background questions: the respondents' primary/secondary area of research (*Q1-1.2*), career stage (*Q2*), work sector (*Q3*), regions of origin and workplace (*Q4-5*), and gender (*Q6*). When asking for region (of origin/workplace), we used the UN geo-scheme, containing 22 sub-regions. The question about gender followed the standards used in the European Social Survey (ESS ERIC, 2024). The responses to these questions are used to describe the sample of the researchers who responded to the survey.

4.2 Intelligence Criteria

As discussed in section 2, there are over 70 definitions of "intelligence" in various fields, many of them overlapping. This would be unfeasible to present in a survey. Hence, we analyzed this list of definitions to identify recurring criteria. We also considered other relevant criteria, based on our own literature review (with sources not covered by the list of Legg and Hutter (2007), from the fields of NLP, ML, AI, Neuroscience, Psychology, and Philosophy). The result was a list of 13 criteria, which we presented to the survey participants as

options to *Q7* (*Which of the criteria are relevant for your use of the term 'intelligence'?*). Our full set of criteria is shown in Figure 1 and Table 1. The latter also lists the key sources for each criterion.

In *Q9*, the respondents were asked to identify the subset of the criteria that they believe to be lacking from current LLM-based systems. In the case of both *Q7* and *Q9* we opted to present this as a binary choice in order to reduce the complexity of the survey. This issues is discussed in the section 7.

4.3 Intelligence of LLMs

Consistently with *Q7* and *Q9*, our *Q8* (*Do you agree that current LLM-based systems are intelligent?*) entails a binary conceptualization of "intelligence", with 4 Likert scale answer options ('strongly agree', 'agree', 'disagree', 'strongly disagree'). An argument can be made that intelligence rather exists on a spectrum (Dennett, 2018). However, the "degrees of intelligence" framing is also problematic in terms of the survey as it presupposes that some entities are more intelligent than others, while it is also possible to have different kinds of cognition not following a hierarchy (De Waal, 2016). Our solution was to present *Q8* as a question specifically targeting the *current* LLM-based systems. We further supplemented it with *Q11*: *Which of the following [entities] would you consider intelligent?* The answer options for this question presented a range of biological and artificial entities.[5]

Using a similar 4-point Likert scale, we also ask the respondents whether they believe that future systems based on a similar technology would satisfy their notion of intelligence (*Q10*).

4.4 Perspectives on the Field

Inspired by Michael et al. (2023), we pose *Q12*: *To what extent is your notion of intelligence shared by other researchers in your field?*. This allows us to estimate to what degree real disagreement on "intelligence" matches the perceived disagreement. We further ask the participants which of the currently used methods for evaluating LLM-based systems they consider applicable to their notion of "intelligence" (*Q13*), and what they consider their research goals (*Q14*).

[5]Biological entities: amoebas, ants, cats, average human adults. Artificial entities: current LLM-based chatbot systems (e.g. ChatGPT), ants, current autonomous LLM-based agents (e.g. based on ChatGPT), current autonomous robotic systems (e.g. self-driving cars), current 'narrow' systems performing a specific task (e.g. chess, protein structure prediction), earlier chatbot system (e.g. customer support bots). The options were randomized, with 'none of the above' as the final option.

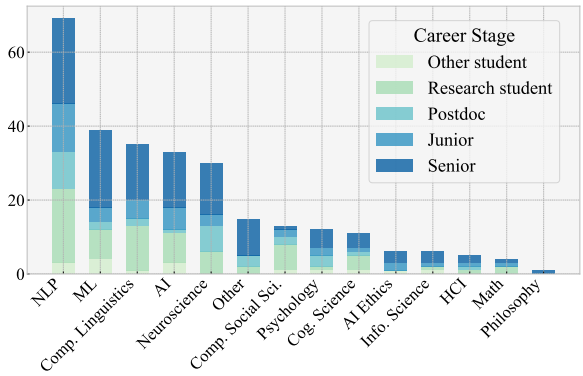


Figure 2: The number of respondents by research area.

4.5 Free-form questions

The survey contains two optional free-form text boxes, giving respondents the opportunity to describe any criteria they find missing among the options given, and to add any additional comments they may have had.

5 Results

5.1 Survey Completion

We received a total of 303 responses which were completed to the end, and 86 partially complete responses. The counts of responses to each question can be found in Appendix A. We found that many participants dropped out of the survey at the start of the section with questions about intelligence (*Q7*), which begins with the criteria of intelligence. We believe this is an interesting finding by itself: **even researchers sufficiently interested in "intelligence" to volunteer for a survey may find it hard to specify what they mean by this term** (even as a multi-choice question with a relatively limited set of commonly mentioned criteria rather than full definitions).

5.2 Demographics

As shown in Figure 2, the survey respondents span a wide variety of research areas, although we received the most responses from NLP, ML, Neuroscience, and Computational Linguistics. The *Other* category includes researchers from fields such as Political Science, Robotics, Economics, and Engineering.

Respondents work in a wide variety of geographic locations, but the majority are employed at organizations in North-Western Europe and North

America. Other areas were selected more often as regions of origin than current employment (see Appendix C for more details). The occupational background of respondents was dominated by 240 respondents working in academia (including students), with an additional 59 respondents from industry, 24 working in government or non-profit organizations and an additional 8 respondents who did not wish to disclose their occupation[6]. The survey is thus skewed towards the academic perspective. $\approx 66\%$ of respondents identified as male, $\approx 25\%$ identified as female, and the remaining 9% either chose not to disclose or selected *Other*.

5.3 Intelligence Criteria

> **RQ1**: *What are key criteria for a system to be considered "intelligent", and to what degree do researchers in different fields, occupations, and career stages agree?*

First we assess whether there are significant group differences in the intelligence criteria selected by the respondents (see the list of criteria in Table 1). We compute the *phi* (ϕ) coefficients using 2x2 contingency tables for each pair of criteria. For binary data the ϕ coefficient is equivalent to the Pearson correlation coefficient. This analysis thus assesses whether respondents are likely to select a criterion A if they have selected a criterion B.

Figure 1 illustrates the criteria selected by respondents (nodes) as well as correlations between criteria (edges). Considering the number of respondents who selected specific criteria of intelligence, **the three top criteria are *generalization* ($\approx 86\%$), *adaptability* ($\approx 83\%$), and *reasoning* ($\approx 83\%$).** *Reasoning* is not strongly correlated with the other two top candidates, whereas *generalization* and *adaptability* show the strongest correlation of all criteria ($\phi = .451$, $p \ll 0.01$).

Figure 3 shows the distribution of criteria selection by occupation. To investigate the connection between selected criteria of intelligence and demographic factors we use Fisher's exact test, to account for the small sample size observed in respondent groups such as HCI, philosophy, and mathematics. We construct a contingency table with selection rates for each group and compare criterion selection rates. The strength of association is quantified via Cramer's V (ϕ_c). **We observe no significant differences in the selection of intelligence criteria when grouping by either occupation, career stage, or primary research area**. The only exception is an effect of career stage on consciousness ($\phi_c = .215$, $p = .006$) and embodiment ($\phi_c = .218$, $p = .01$), which is driven by low selection rate of these criteria by post-docs. Likewise, we observe an effect for environment interaction ($\phi_c = .225$, $p = .003$), which is driven by its low selection rate by students. These results indicate an overall coherence in perspectives on "intelligence" that contrasts with career stage effect on the perceived intelligence of current LLM-based systems (subsection 5.4).

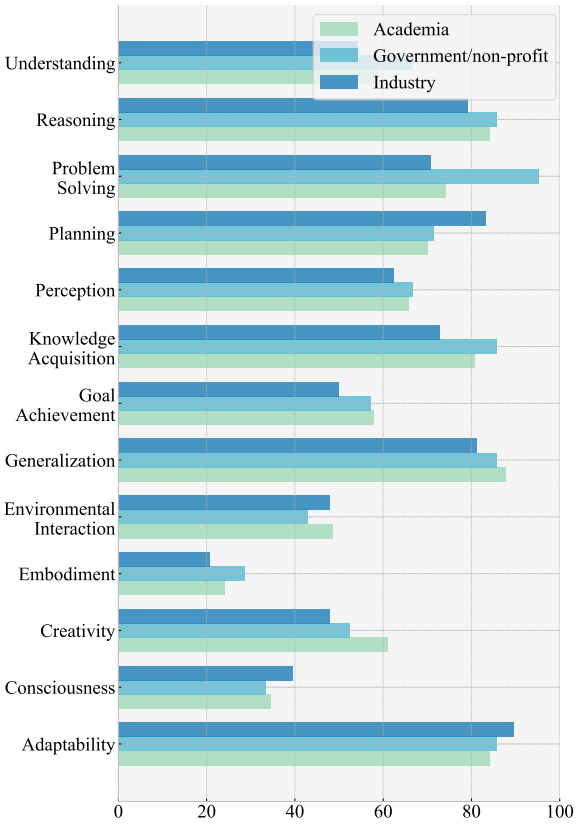


Figure 3: The percentage of respondents (x-axis) who selected a given intelligence criteria (y-axis) by primary occupation (color). *Other* is removed for readability.

We additionally asked which of the 13 criteria are lacking in the current LLM-based systems. Table 2 shows the percentage of respondents who selected various criteria as important and who consider them lacking in the current systems. For most criteria, more respondents listed them as important for their notion of "intelligence" than those who listed them as lacking in the current systems, which suggests that researchers are hesitant to state definitively what current systems cannot do.

[6]The question is multi-select, as it is possible to be employed in both academia and industry simultaneously.

Criterion	Important	Lacking
% Important overall > % Lacking in LLMs		
Reasoning	83%	59%
Generalization	86%	54%
Adaptability	83%	55%
Planning	70%	58%
Knowledge acquisition	79%	44%
Perception	63%	47%
Problem solving	74%	32%
Creativity	56%	50%
Understanding	57%	36%
Env. interaction	45%	43%
Goal Achievement	55%	32%
% Lacking in LLMs > % Important overall		
Consciousness	35%	63%
Embodiment	23%	53%

Table 2: Percentage of respondents who selected the listed criteria as important for their notion of "intelligence", and who consider them lacking in the current LLM-based systems.

For the top 3 criteria of intelligence selected by the majority of respondents (*generalization, adaptability & reasoning*), all of them are considered to be lacking in the current systems by 54-59% of respondents. On the other end of the spectrum, only 32-36% of respondents stated that the current systems are lacking in *goal achievement, problem solving, & understanding*. The latter is surprising, given that many respondents come from NLP & Computational Linguistics communities, where numerous studies and position papers highlighted the model reliance on statistical patterns as evidence against "understanding" (Bender and Koller, 2020; Ray Choudhury et al., 2022; Mitchell and Krakauer, 2023; Wu et al., 2024, among others). This number indicates a change since 2022, when a different survey found an even split in the participants asked whether text-only generative models could be said to "understand natural language in some non-trivial sense" (Michael et al., 2023).

The only criteria which are considered lacking in LLMs by a higher percentage of respondents than those who consider it overall relevant for "intelligence" are *consciousness & embodiment*. For the latter it is particularly hard to make a case that it is present in the current systems like ChatGPT, yet, surprisingly, 47% do not select it as lacking.

35 of the 303 respondents described additional criteria. They highlighted, in particular, social and emotional traits, and the fact that intelligence should be understood as a spectrum rather than a binary. Self-awareness and being alive or conscious was also a recurrent comment, and a few respondents emphasized abilities in knowledge-acquisition and more abstract generalization. Respondents additionally emphasized the difficulty in defining some of the criteria we use to pinpoint their notion of intelligence. A sample of free-text comments is available in Appendix E.

Strongly agree	Agree	Disagree	Strongly disagree
"Current LLM-based systems are intelligent"			
3%	27%	40%	**31%**
"Future systems will be intelligent"*			
7%	33%	44%	**16%**

Table 3: Difference in responses on whether current LLM-based systems or future systems (*based on a similar technology) are intelligent.

5.4 Intelligence of LLMs

RQ2: *What is the perception of the "intelligence" of current LLM-based systems such as ChatGPT?*

Table 3 shows that **the majority of the survey respondents do not consider the current systems intelligent** ($\approx 71\%$). When they are asked about future systems (based on a similar technology), skepticism remains the majority position ($\approx 60\%$), but there are respondents moving both from 'agree' to 'strongly agree', and from 'strongly disagree' to 'disagree'. The increase in strong agreement indicates a stronger belief that future LLMs will be intelligent, but the decrease in strong disagreement could also signal the unwillingness to strongly commit to a prediction of the future developments. We visualize these trends for specific groups as Sankey diagrams. Figure 4 presents these results for the senior researchers, and similar diagrams for other career stages are available in Appendix D.

Next, we examine the possible effect of demographic factors on the beliefs of intelligence in the current LLM-based systems.

Career Stage Effect. We observe a significant effect of career stage ($\phi_c = 0.164$, $p = 0.033$). Early-career researchers (postdocs and students) are more likely to consider the current systems intelligent ($\approx 48\%$ of postdocs agree or strongly agree, for students that number is $\approx 56\%$). Senior researchers demonstrate greater skepticism, with $\approx 77\%$ disagreeing or strongly disagreeing.

Research Area Effects. We observe a significant effect of research area ($\phi_c = 0.270$, $p = 0.001$). Researchers are generally skeptical of LLM intelligence, with the majority of respondents either dis-

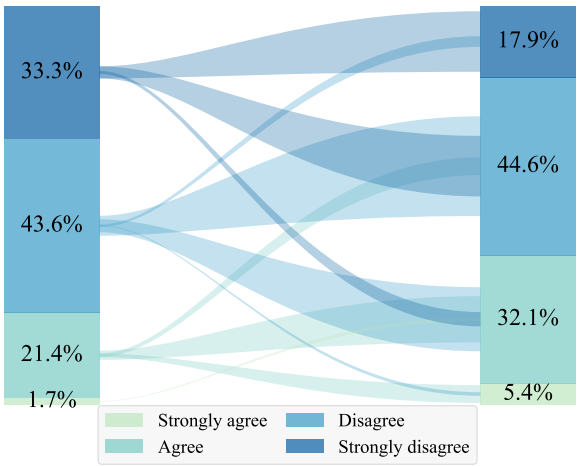


Figure 4: Sankey diagram illustrating the flow of beliefs concerning intelligence in current and future LLM based systems for senior researchers. See Appendix D for additional groups.

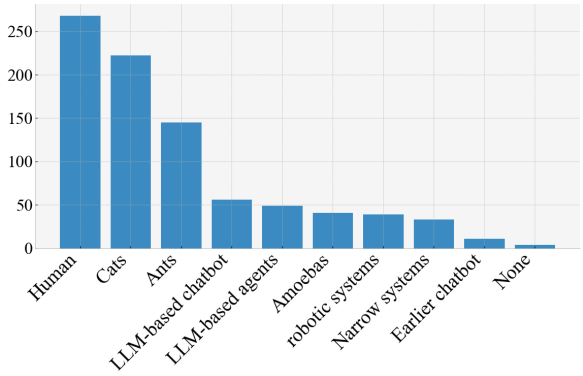


Figure 5: Number of respondent selecting an entity when asked whether the entity should be considered intelligent.

agreeing or strongly disagreeing in every research area except for HCI (60% agree that LLM-based systems are intelligent). However, we have relatively few responses from HCI researchers (8), and do not consider this result reliable. Following HCI, researchers from Cognitive Science and CSS are the most positive regarding LLM intelligence, both with $\approx 45\%$ positive responses. Interestingly, the most skeptical groups are the respondents who selected "AI" or "Computational Linguistics". The vast majority in both groups did not agree that "current systems are intelligent" ($\approx 73\%$ and $\approx 86\%$ respectively).

We conducted a similar analysis with respect to views on the intelligence of future systems, and found no significant differences among demographic groups.

Are LLMs as intelligent as other entities? As discussed in subsection 4.3, we opted for a binary perspective on whether the current LLM-based systems are intelligent. However, we also provided the respondents with the opportunity to rate a wide range of biological and artificial systems as intelligent or not. The distribution of their answers is presented in Figure 5. Almost all respondents stated that humans are intelligent. All biological entities except amoebas have more respondents rating them as intelligent, than all artificial systems. Within the latter, the LLM-based systems are selected as intelligent more often than others (though the difference is not as significant as between the biological entities). In the order of number of researchers rating various entities as intelligent, **the current LLM-based systems rank between ants and amoebas.** Interestingly, the agentic systems have slightly fewer supporters than the non-agentic ones.

5.5 Perspectives on the Field

> **RQ3**: *What role, if any, does ones research agenda play in the notion of "intelligence"?*

Research Agenda. Following the same methodology as used in subsection 5.3, we assess whether the research agenda of the respondents influences their selection of intelligence criteria and their perception of intelligence of LLM-based systems. We observe no significant differences with regards to the former. For the latter, we find that the researchers focusing on the interaction of tech and society are more likely to 'strongly disagree' with the statement that these systems are intelligent (for the current systems $\phi_c = .248$, $p = .001$, for the future systems $\phi_c = .187$, $p = .032$). Likewise, **the researchers who aim at creating intelligent systems are more likely to consider LLM-based systems as intelligent** (for the current systems $\phi_c = .208$, $p = .004$, for future systems $\phi_c = .216$, $p = .01$). This is an interesting observation, given the above finding that the respondents identifying their primary field of research as "AI" are among the most skeptical about the current systems. This suggests a dissociation between "AI" as a research field and "intelligent technology" as a research goal.

Perception of what other researchers believe. We find that $\approx 45\%$ **of respondents believe that**

Research Goal	Selection rate
Adding to Scientific Knowledge	72.5%
Creating Intelligent Technology	16.2%
Creating Practically Useful Technology	58.1%
Discussing Tech / Societal Interaction	21.8%
Other	4.6%

Table 4: Overview of respondent research goals. Respondents were allowed to select multiple options.

more (or most) of their peers would agree than disagree with their notion of "intelligence". A further $\approx 35\%$ are unsure about the perspective of their peers. A minority of the respondents believe that their notion of "intelligence" is in the minority ($\approx 20\%$).

Tests of Intelligence. The vast majority of respondents (62%) do not believe that we have a test adequately measuring "intelligence". The respondents generally rejected all listed options (including composite benchmarks, Turing test, IQ and professional knowledge tests). The option that had the most support was generalization tests (37%).

Feedback on the survey and additional comments. We received 33 unstructured comments. A part of them was feedback on the survey. Some respondents commented that the question options were too forced, that intelligence is a spectrum (see subsection 4.3), and a few argued that intelligence is not a useful concept for assessing neither humans nor systems.

We also received more general comments towards the discussion of intelligence, including different perspectives on how we could understand this term as something an entity can *have* or behave according to, rather than something an entity can *be* (see Appendix E for a sample of free-text comments).

6 Discussion

This survey provides concrete numbers to a core source of confusion in the field: what we mean when we use the term "intelligence". While most papers using this term do not provide a definition, and many researchers struggle to specify what they mean (subsection 5.1), we do observe a degree of coherence. The majority of respondents ($> 80\%$), regardless of their career stage, primary research area and occupation, agree that this notion relies on the criteria of *generalization*, *adaptability* and *reasoning* (**RQ1**). The high agreement on these criteria offers the community an opportunity to make the discussion of "intelligence" more specific, and encourage evaluation that focuses on them. This would also be consistent with the fact that generalization tests, such as the ARC challenge (Chollet, 2019), was the most selected among respondents ($\approx 37\%$).

Given that there is some consistency to the core criteria of "intelligence", we would expect significant agreement on whether this term applies to the current LLM-based systems (**RQ2**). Indeed, this is what we find: the majority of respondents are skeptical about both current ($\approx 71\%$) and future systems based on a similar technology ($\approx 60\%$) (subsection 5.4). This skepticism stems in part from the limitations of current systems to generalize and reason (Table 2). Given that most respondents consider these criteria core to the notion of "intelligence", they would be expected to be skeptical of definitions that do not rely on these criteria[7].

Finally, the results of this survey provide context for the recent debate on the relation between the fields of "NLP" and "AI" (Bender, 2024; Goldberg, 2024; Mortensen, 2024). The majority of survey respondents do not see their goal as creating intelligent systems, focusing instead on advancing the scientific knowledge. However, we found an interesting correlation between research goals and beliefs regarding the intelligence of current systems (**RQ3**). The respondents who do aim to create intelligent systems were more likely to attribute intelligence to both the current and future LLM-based systems.

7 Conclusion

The term "intelligence" is often used, but rarely defined in the current research on Large Language Models. Despite that, we find a high degree of consensus across research fields that *generalization*, *adaptability*, and *reasoning* are key to this notion. The majority of the survey respondents are skeptical of applying this term to the current and future systems based on LLMs. We find that the senior researchers tend to be more skeptical, and those respondents whose research agenda aligns with creating intelligent systems are more likely to attribute intelligence to current and future systems. Our results highlight a discrepancy between how researchers conceptualize intelligence and how this term is used in public discourse and marketing.

[7] Notably, OpenAI was reported to define 'AGI' in terms of profit that this technology can yield (Zeff, 2024).

Limitations

Comprehensiveness. Defining "intelligence" is in many ways an impossible task even in a long-form text. This study covered only a small subset of criteria that can be found in existing definitions of "intelligence" (and for which the scope of this study would not allow for a full literature review). As discussed in subsection 4.2, we aimed to present at least the core criteria, and most of the criteria we chose are commonly mentioned in the current literature on LLM-based systems.

Intelligence as a binary attribute. To keep the survey question complexity manageable, we had to simplify many questions, including the binary choice of different criteria as either present or not present in different systems. Given the difficulty that our respondents had even with this design (section 5), we recommend that future work on this issue would consider either more focused surveys that ask fewer questions in greater detail, or a different methodology.

Our Likert scale questions about applicability of the term "intelligence" to the current/future LLM-based systems could present it as a variable on a spectrum, but we opted for the binary presentation for the reasons outlined in subsection 4.3.

Assumption of coherent notion of "intelligence". Our survey targets the use of the term "intelligence" by researchers, on the assumption that for each individual there is a coherent notion underlying their use of this term. It is possible that this assumption is false, and researchers use this term differently for different entities (cf the proposal to consider LLMs as "a new type of intelligence" by McCoy et al. (2024)). This possibility necessitates further research.

Definitions of intelligence criteria. Many of the criteria we listed, especially "reasoning" and "understanding", are themselves lacking clear definitions, and merit similar surveys. Hence, while we can conclude that respondent views on intelligence criteria are highly coherent, we cannot be certain that each of the 13 intelligence criteria have the same definition for every respondent.

Representativeness. The survey respondents necessarily present only a sample of the researcher population, which poses questions of its representativeness. As discussed in subsection 5.2, our sample is skewed towards academics from the "western" world, and may not represent the views of researchers in other regions and organization types. To mitigate that, we recommend that future work should identify and reach out to distribution channels beyond the mailing lists and social media based in the West.

Selection bias. We acknowledge the possibility that since the respondents sample is affected by self-selection bias, drawing the attention of the researchers who are more interested in "intelligence" than average. Given that, our result of 16.2% respondents who view intelligent systems as a research goal may be an overestimate.

Ethical Considerations

Broader impact This study maps the criteria that are believed by researchers to be important for building intelligent systems. The survey merely provides a snapshot of the sentiment in 'AI' related research communities, but we hope that our results will help frame the discussion on what intelligence is and how to develop intelligent systems. It is further the hope that our survey can serve as a useful point of comparison for future studies in a field that is evolving at a rapid pace.

Personal and sensitive information All responses were completely anonymous and thus designed to not solicit any personally identifiable information or fine-grained demographic information about participants. Since no personal or contact information was collected, it was not possible to provide the respondents with the option to withdraw or alter their responses after the completion of the survey.

Due to low response-rates for certain subgroups, the publicly available dataset is curated to remove some demographic information. In particular, we remove the gender data, and merge the subdivisions of geographical regions, so that e.g. both Eastern and Middle Africa are coded as "Africa".).

Data and code availability All code used to perform statistical tests and report results, as well as survey data and full questionnaire is included in the supplementary materials, and will be publicly available under CC BY-NC 4.0 license upon the publication of this study.

Competing interests The authors have no relevant financial or non-financial interests to disclose.

Ethics Statement The study was approved by the Research Ethics Committee at the authors' institution.

References

J.S. Albus. 1991. Outline for a theory of intelligence. 21(3):473–509.

Anthropic. Core Views on AI Safety: When, Why, What, and How.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Emily M. Bender. 2024. ACL Is Not an AI Conference.

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *Preprint*, arXiv:2303.12712.

Tyler A. Chang and Benjamin K. Bergen. 2023. Language Model Behavior: A Comprehensive Survey.

Junzhe Chen, Xuming Hu, Shuodi Liu, Shiyu Huang, Wei-Wei Tu, Zhaofeng He, and Lijie Wen. 2024. Llmarena: Assessing capabilities of large language models in dynamic multi-agent environments. *Preprint*, arXiv:2402.16499.

François Chollet. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. *Preprint*, arXiv:2110.14168.

Frans De Waal. 2016. *Are we smart enough to know how smart animals are?* W. W. Norton & Company.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2024. Investigating data contamination in modern benchmarks for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8706–8719, Mexico City, Mexico. Association for Computational Linguistics.

D. C. Dennett. 2018. *From Bacteria to Bach and Back: The Evolution of Minds*, fist published as a norton paperback edition. W. W. Norton & Company.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. 2024. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12039–12050, Bangkok, Thailand. Association for Computational Linguistics.

ESS ERIC. 2024. European Social Survey (ESS), Round 11 - 2023.

Howard Gardner. 1983. *Frames of Mind: The Theory of Multiple Intelligences*. Basic Books, New York.

Matt Gardner, William Merrill, Jesse Dodge, Matthew Peters, Alexis Ross, Sameer Singh, and Noah A. Smith. 2021. Competency problems: On finding and removing artifacts in language data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1801–1813, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2020. Datasheets for Datasets. *arXiv:1803.09010 [cs]*.

Gilles E. Gignac and Eva T. Szodorai. 2024. Defining intelligence: Bridging the gap between human and artificial perspectives. 104:101832.

Yoav Goldberg. 2024. ACL is not an AI Conference (?).

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. *Preprint*, arXiv:2009.03300.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5084, Singapore. Association for Computational Linguistics.

Joohee Kim and Il Im. 2023. Anthropomorphic response: Understanding interactions between humans and artificial intelligence agents. 139:107512.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253.

Yann LeCun. 2022. A Path Towards Autonomous Machine Intelligence.

Shane Legg and Marcus Hutter. A Formal Measure of Machine Intelligence. *Preprint*, arXiv:cs/0605024.

Shane Legg and Marcus Hutter. 2007. A Collection of Definitions of Intelligence. *Preprint*, arXiv:0706.3639.

Jiaang Li, Yova Kementchedjhieva, Constanza Fierro, and Anders Søgaard. 2024. Do vision and language models share concepts? a vector space alignment study. *Transactions of the Association for Computational Linguistics*, 12:1232–1249.

Gary F Marcus. 2003. *The algebraic mind: Integrating connectionism and cognitive science*. MIT press.

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D. Hardy, and Thomas L. Griffiths. 2024. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Drew McDermott. 1976. Artificial intelligence meets natural stupidity. (57):4–9.

William Merrill, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. 2021. Provable limitations of acquiring meaning from ungrounded form: What will future language models understand? *Transactions of the Association for Computational Linguistics*, 9:1047–1060.

Julian Michael, Ari Holtzman, Alicia Parrish, Aaron Mueller, Alex Wang, Angelica Chen, Divyam Madaan, Nikita Nangia, Richard Yuanzhe Pang, Jason Phang, and Samuel R. Bowman. 2023. What do NLP researchers believe? results of the NLP community metasurvey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16334–16368, Toronto, Canada. Association for Computational Linguistics.

Margaret Mitchell, Alexandra Sasha Luccioni, Nathan Lambert, Marissa Gerchick, Angelina McMillan-Major, Ezinwanne Ozoani, Nazneen Rajani, Tristan Thrush, Yacine Jernite, and Douwe Kiela. 2023. Measuring data. *Preprint*, arXiv:2212.05129.

Melanie Mitchell. 2021. Why AI is Harder Than We Think. *Preprint*, arXiv:2104.12871.

Melanie Mitchell and David C. Krakauer. 2023. The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120.

David R. Mortensen. 2024. Is ACL an AI (or NLP or CL) Conference?

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. *Preprint*, arXiv:2301.05217.

Catherine Olsson, Nelson Elhage, and Neel Nanda. 2022. In-context Learning and Induction Heads.

OpenAI. OpenAI Platform.

OpenAI. 2022. Introducing ChatGPT.

R Pfeifer. 2006. How the body shapes the way we think: A new view of intelligence.

Rolf Pfeifer and Christian Scheier. 2001. *Understanding intelligence*. MIT press.

Jean Piaget. 1947. *La Psychologie de l'Intelligence*. Armand Colin, Paris.

John Piaget. 1952. The origins of intelligence in children. *International University*.

Inioluwa Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the Everything in the Whole Wide World Benchmark.

Sagnik Ray Choudhury, Anna Rogers, and Isabelle Augenstein. 2022. Machine Reading, Fast and Slow: When Do Models "Understand" Language? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 78–93. International Committee on Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Anna Rogers and Sasha Luccioni. 2024. Position: Key Claims in LLM Research Have a Long Tail of Footnotes. In *Forty-First International Conference on Machine Learning*.

Naomi Saphra and Sarah Wiegreffe. 2024. Mechanistic? In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 480–498, Miami, Florida, US. Association for Computational Linguistics.

David Schlangen. 2021. Targeting the benchmark: On methodology in current natural language processing research. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 670–674, Online. Association for Computational Linguistics.

Linda Smith and Michael Gasser. 2005. The development of embodied cognition: Six lessons from babies. *Artificial life*, 11(1-2):13–29.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 431 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*. Featured Certification.

Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. 2011. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285.

Alan M. Turing. 1950. I.—COMPUTING MACHINERY AND INTELLIGENCE. LIX(236):433–460.

Marcel Van Gerven. 2017. Computational foundations of natural intelligence. *Frontiers in computational neuroscience*, 11:299674.

David Vilares and Carlos Gómez-Rodríguez. 2019. HEAD-QA: A Healthcare Dataset for Complex Reasoning. *Preprint*, arXiv:1906.04701.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *Preprint*, arXiv:1905.00537.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355. Association for Computational Linguistics.

Joseph Weizenbaum. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. 9(1):36–45.

Geraint A Wiggins. 2020. Creativity, information, and consciousness: the information dynamics of thinking. *Physics of life reviews*, 34:1–39.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1819–1862, Mexico City, Mexico. Association for Computational Linguistics.

Maxwell Zeff. 2024. Microsoft and OpenAI have a financial definition of AGI: Report.

A Appendix: Survey Questions

Table 5 provides a full list of survey questions together with answer options. The survey is also available as a form in the supplementary materials.

B Appendix: Mailing lists

Table 6 provides a full list of mailing lists to which the survey was sent.

C Appendix: Geographic Distribution of Respondents

Figure 6 provides an overview of the geographical distribution of respondents.

D Appendix: Sankey Diagrams: Belief Development

Figure 7 showcases the beliefs of subgroups the respondents regarding the intelligence of LLM-based systems.

E Appendix: Free-text Comments

Table 7 and Table 8 provide a sample of comments from survey respondents.

E.1 Selected Comments on Criteria of Intelligence

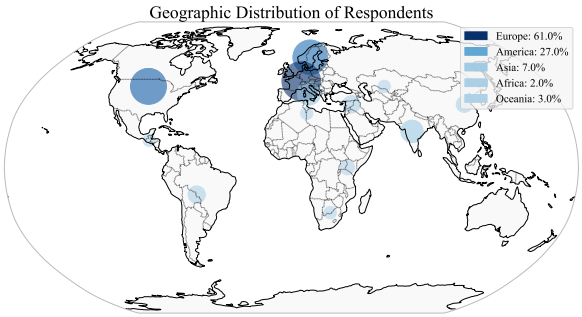
E.2 Selected General Comments on the Survey

ID	Question	no. of responses
Q1	What area of research do you consider your primary area?	303
Q1.2	If you also work in another area please select that: Answer options for Q1-1.2: *AI Ethics/Governance, Artificial Intelligence, Computational Linguistics, Computational Social Science, Cognitive Science, Human-Computer Interaction, Information Science, Machine Learning, Mathematics, Natural Language Processing, Neuroscience, Philosophy, Psycholinguistics, Psychology (cognitive, developmental...), Other*	256
Q2	What is your current career stage? Answer options: *Senior (faculty or industry), Junior (faculty or industry), Postdoc, Research student (PhD, MPhil), Other student (Master, Bachelor), Other, Prefer not to say*	303
Q3	In what sector do you currently work? Answer options: *Academia, Industry, Government/non-profit, Prefer not to say*	303
Q4	What is your region of origin?	303
Q5	In what region is your primary place of work? Answer options for Q4-5: 22 regions from the UN geoscheme (https://en.wikipedia.org/wiki/United_Nations_geoscheme)	303
Q6	Which of the [gender] options best describe you? Answer options: *A woman, A man, Other, Prefer not to say, Don't know*	303
Q7	Which of the criteria are relevant for your use of the term 'intelligence'? Answer options: *Interaction with Environment (physical or virtual), Embodiment (being situated in a physical environment), Perception (extracting and acting upon useful information from the environment), Knowledge acquisition (learning, understanding or gaining knowledge and skills), Problem solving (in a familiar domain), Goal Achievement (accomplishing defined objectives and optimizing for performance on specific tasks), Reasoning (logical inference - deductive, abductive etc.), Planning (anticipating future events and organizing actions based on a deliberate strategy), Adaptability (making sense of new environments and/or handling novel tasks), Generalization (successfully handling new types of data and situations), Creativity (in your definition of this term), Consciousness (in your definition of this term), Language understanding (in your definition of this term)*	302
Q8	Do you agree that current LLM-based system are intelligent? Answer options: Strongly disagree, disagree, agree, strongly agree	300
Q9	Which of the criteria of intelligence are lacking in the current LLM based systems (such as ChatGPT)? Answer options: same as for Q7	298
Q10	If current LLM-based systems do not satisfy your criteria for intelligence, do you agree that future systems based on similar technology will? Answer options: *Strongly disagree, disagree, agree, strongly agree*	291
Q11	Which of the following [entities] would you consider intelligent? Answer options: *Average human adults , Cats, Ants, Amoebas, Current 'narrow' systems performing a specific task (e.g. chess, protein structure prediction), Current autonomous robotic systems (e.g. self-driving cars), Earlier chatbot system (e.g. customer support bots), Current LLM-based chatbot systems (e.g. ChatGPT), Current autonomous LLM-based agents (e.g. based on ChatGPT), None of the above*	299
Q12	To what extent, is your notion of intelligence shared by other researchers in your field? Answer options: *Most would agree with me, More would agree than disagree, More would disagree than agree, Most would disagree with me, Unsure*	298
Q13	Which of the following tests, if any, do you believe measure intelligence Answer options: *Human preference ranking of outputs of different models, Tests for measuring human intelligence, such as standardized IQ tests , Tests for measuring human professional knowledge, e.g. SAT or medical exam questions, Average scores on composite benchmarks for LLMs such as MMLU or BIG-bench, The Turing test, Generalization tests (focusing on differences between training and test distributions), Other (please specify), I do not believe that we currently have such a test*	299
Q14	Which of the following options are the best description of your research goals? Answer options: *Creating technology that qualifies for my notion of 'intelligence', Creating technology that is practically useful, Documenting/critiquing the interaction between technology and the society/ecosphere, Adding to the knowledge/scientific understanding of the phenomenon I study (e.g. cognition, language, technology, society etc.), Other*	296

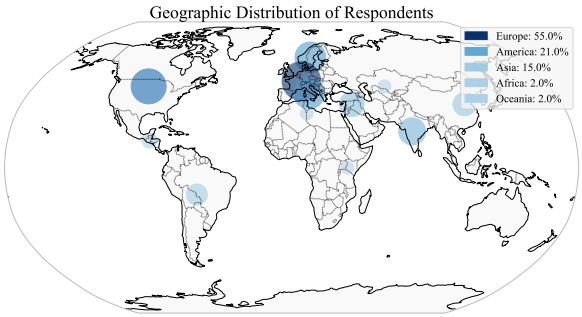
Table 5: Overview of questions as phrased in the survey and number of respondents who completed each question (out of 303 respondents who made it to the end of the survey). The questions with multi-choice answer options were presented in the randomized order, with "Other" position fixed at the end.

Mailing List	Link
ML News	https://groups.google.com/g/ml-news
Corpora List	https://list.elra.info/mailman3/hyperkitty/list/corpora@list.elra.info/
EUcog News	https://groups.google.com/g/eucog-general-news
Sys. Neuroscience	https://groups.google.com/g/systems-neuroscience
Neural Ensemble	https://groups.google.com/g/neuralensemble
Connectionist	https://mailman.srv.cs.cmu.edu/mailman/listinfo/connectionists
ACL portal	https://www.aclweb.org/portal/

Table 6: Overview of mailing lists used to distribute the survey.

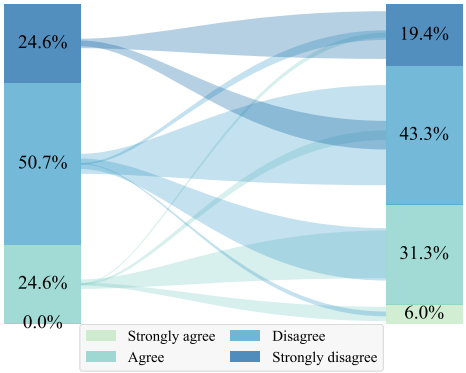


(a) Geographical distribution of respondents based on their region of work.

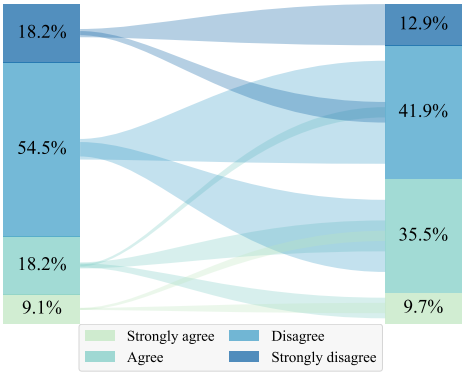


(b) Geographical distribution of respondents based on their region of origin.

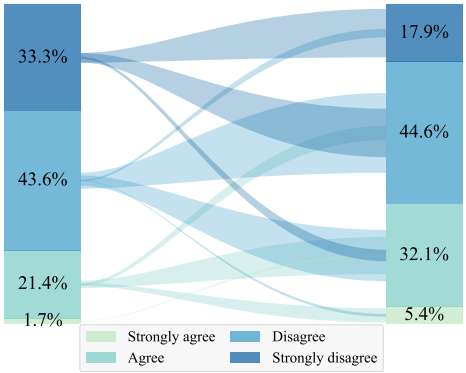
Figure 6: Geographical distribution of survey respondents. Most respondents are from the "western" world, although more respondents are originally from e.g. Asia and Africa than currently work in those regions.



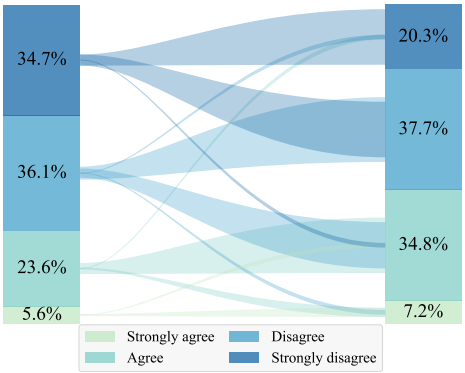
(a) Sankey diagram showing changing beliefs for "NLP" researchers.



(b) Sankey diagram showing changing beliefs for "AI" researchers.



(c) Sankey diagram showing changing beliefs for senior researchers.



(d) Sankey diagram showing changing beliefs for student researchers (PhD level).

Figure 7: Sankey diagrams showing the different beliefs of researchers regarding the intelligence of current LLM-based systems (on the left) and future systems based on similar technology (on the right).

Comment ID	Text
C1	Emotional capacity, understanding and feeling a sense of qualia, and having some sense of humanness/mammalness that invokes/evokes responses similar to carbon based life-forms. I'm keeping this to Carbon based emulation because other loose forms of consciousness becomes very unbounded and we don't have sufficient tools or language yet to describe or understand it.
C2	Abstraction: being able to make a new concept that abstracts the core similarities of the underlying group. If using LLMs as an example, then imagine a vector that is able to be decoded into a new never before seen token. This is to be differentiated from generalisation.
C3	Many of these are relevant to "intelligence", but they each sit on a spectrum. So determining "intelligence" isn't generally going to be a box-ticking exercise.
C4	The notion of "intelligence" itself is deeply problematic, with origins in eugenics and racism. The idea that people (and other beings) can be ranked according to such a property is abhorrent. The survey did not let me choose none of them, so I chose some, but under duress.
C5	I've answered this under what I think of as "weak intelligence," where consciousness isn't required but some abstract reasoning is required.

Table 7: Selected free-text comments from the survey participants that focus on the criteria and definitions of intelligence.

Comment ID	Text
C1	Some of the questions forced me to an oversimplification. I understand that this is, to some extent, necessary to produce immediate answers, and you cannot expect a treatise on the topic from every participant. However, adding more "unsure" or "it depends" options could have helped.
C2	Because each dimension of intelligence is a matter of degree, I can't answer yes/no questions about whether something is intelligent despite your efforts to force me to do so. Hence, most of the questions on this page are poorly formulated. LLMs, for example, are more broadly knowledgeable and have better language comprehension than any previous system. They are more intelligent along these dimensions than other systems. But there is no threshold that separates "intelligent" from "unintelligent".
C3	To me, intelligence is a spectrum – humans and other animals can be more or less intelligent, and it is difficult to draw a line. I think this spectrum is, by definition, centered around *human* intelligence, and so it is easier to place something on the spectrum if it behaves more like humans. LLMs do not behave in a very human-like manner, so it is difficult to measure their intelligence.
C4	I actually think that "intelligence" is not a useful concept to assess human and animal cognition, not is it a useful concept to assess AI systems. Human and animal cognition have quite specific properties that cover a range of needs (orienting action at objects/other beings, social competence, learning from experience, etc). These can be assessed in ways known from psychology and neuroscience. Artificial agents typically have much narrower ranges of skills or properties. It may be misleading to measure and compare these to organisms through a single concept.
C5	Systems that pass the Turing test may not actually BE intelligent, but maybe the distinction between intelligence and not becomes meaningless at a certain point? Awareness, empathy and altruism are key elements in (my interpretation of) intelligence, and I seriously doubt that we will see all three traits combined in a silicon 'mind' anytime soon.

Table 8: Selected free-text comments from the respondents on the general considerations on the criteria, on how to think about "intelligence", and general concerns regarding the discussion.