BenCSSmark: Making the Social Sciences Count in LLM Research

Arnault Chatelain[1], **Étienne Ollion**[1], **Qianwen Guan**[2], **Diandra Fabre**[3], **Lorraine Goeuriot**[3], **Emile Chapuis**[4], **Abdelkrim Beloued**[4], **Marie Candito**[2], **Nicolas Hervé**[4], **Didier Schwab**[3]

[1] *CREST (École Polytechnique, ENSAE, CNRS), 5 avenue Le Chatelier, 91120 Palaiseau, France*

[2] *LLF (Université Paris Cité and CNRS), UFRL Olympe de Gouges, 13 place Paul Ricoeur, 75013 Paris, France*

[3] *Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France*

[4] *INA (Institut National de l'Audiovisuel), 4 Avenue de l'Europe, 94366 Bry-sur-Marne, France*

arnault.chatelain@ensae.fr, etienne.ollion@polytechnique.edu,

{qianwen.guan, marie.candito}@u-paris.fr

{echapuis, abeloued, nherve}@ina.fr,

{diandra.fabre, lorraine.goeuriot, didier.schwab}@univ-grenoble-alpes.fr

Abstract

This position paper argues that the under-representation of social science tasks in contemporary LLM benchmarks limits advances in both LLM evaluation and social scientific inquiry. Benchmarks — standardized tools for assessing computational systems — are pivotal in the development of artificial intelligence (AI), including large language models (LLMs). Benchmarks do more than measure progress — they actively structure it, shaping reputations, research agendas, and commercial outcomes. Despite this central role, the social sciences are largely absent from mainstream evaluation frameworks, even though scholars in these fields generate dozens of rigorously annotated, context-sensitive datasets each year. Integrating this work into benchmark design could significantly improve the generalization and robustness of AI models. In turn, models trained on social scientific tasks would likely yield better performance on classic and contemporary tasks in disciplines as diverse as history, sociology, political science or economics. This is all the more pressing as these disciplines are quickly turning to LLMs for assistance. To address this gap, we introduce BenCSSmark, a benchmark composed of datasets annotated by computational social scientists. By integrating social scientific perspectives into benchmarking, BenCSSmark seeks to promote more robust, transparent, and socially relevant AI systems and to foster efficient collaboration.

Keywords: LLMs, Benchmarks, Social Sciences, Computational Social Sciences, Evaluation, Dataset, Annotation, Perspectivism

1. Introduction

Benchmarks — tools designed to evaluate the performance of computational systems — have played a central role in the development of artificial intelligence (AI) (Koch et al., 2021; Raji et al., 2021 ; for a historical perspective, see Orr and Kang, 2024). By providing standardized measures of performance, benchmarks make it possible to compare a system's quality with that of its predecessors or competitors. When it comes to large language models (LLMs), benchmarking has become an ubiquitous practice, with certain names now familiar to virtually all practitioners in the field.

Benchmarks are so central that their outcome often determines the success or failure of new models. Today, they shape both the perceived credibility and the market visibility of any LLM. But benchmarks do not merely measure progress; they actively guide it (Jaton, 2021). They steer research priorities toward incremental improvements in benchmark scores (Luitse et al., 2025), and they delineate the scope of what counts as success in the field (Crawford and Paglen, 2021). Their influence has become so pervasive that some scholars now refer to them as a "lottery" (Dehghani et al., 2021), underscoring the extent to which research trajectories and reputations hinge on performance within a specific evaluative framework.

As their importance grew, the number of available benchmarks has dramatically increased (for an extensive review, see Ni et al., 2025 and Liu et al., 2024, Section 5). Benchmarks are now used to evaluate and select large language models on a wide variety of tasks and capabilities. Many focus on traditional linguistic tasks in natural language processing, such as paraphrasing (Zhang et al., 2019; Yang et al., 2019), named entity recognition (Mayhew et al., 2024), summarization (Hasan et al., 2021), or translation (NLLB Team et al., 2022) with some aggregating tasks (e.g. natural language inference, coreference resolution, disambiguation) into general language understanding benchmarks (Wang et al., 2018, 2019). Increasingly, benchmarks have also started to assess models' reasoning, their general knowledge, their factual consistency, and their alignment with human values (Hendrycks et al., 2020, 2021a; Zhao et al., 2024), sometimes as part of a multitask effort testing several capabilities at the same time (Srivastava et al., 2023; Suzgun et al., 2022).

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 848–859
11–16 May 2026.

Beyond these general-purpose evaluations, increasingly specialized benchmarks have appeared. Some test mathematical reasoning (e.g., GSM8K, (Cobbe et al., 2021), MATH (Hendrycks et al., 2021b)), while others assess code generation and program synthesis (e.g., HumanEval (Chen et al., 2021), MBPP (Austin et al., 2021)). Domain-specific benchmarks now exist across diverse fields, including medicine (e.g., PubMedQA, MedMCQA), law (e.g., LegalBench, CaseHOLD), finance (e.g., FinQA, FPB), and education (e.g., EduBench). This proliferation reflects both the diversification of model applications and the desire to capture domain-sensitive competencies that generic benchmarks often overlook.

Despite this growing diversity, very few benchmarks are specifically designed for tasks relevant to the social sciences. This set of disciplines which includes, among others sociology, economics, political science, history, geography, anthropology, and demography, is also rarely represented in general benchmarks reviews. Popular leaderboards (e.g. Open LLM Leaderboard (Hugging Face, 2025), MTEB Leaderboard (Muennighoff et al., 2023)) only partially incorporate tasks explicitly labeled as social science tasks. And when they are included, this representation is limited.

This absence is unfortunate for two reasons. First, LLMs are increasingly used in the social sciences, and they are often produced with detailed guidelines and test sets. The quality of the annotation is usually excellent, as the tasks tend to be annotated by domain experts, rather than by standard annotators whose work can be less consistent even in mainstream benchmarks (Klie et al., 2024). Second, the tasks in the social sciences are original, and they embed cultural differences and diverse viewpoints that contemporary benchmarks lack.

Yet these available, high-quality resources often remain invisible — uncatalogued, dispersed, and therefore effectively non-existent from a computer scientist's perspective. Because they are not consolidated into available datasets, they do not appear in mainstream suites. Models are subsequently seldom evaluated on them. This probably explains why even state-of-the-art models struggle to meet the needs of social science research, as evidenced by the wide variance in performance on ostensibly similar tasks (Ollion et al., 2023).

As part of Pantagruel, a scientific project dedicated to producing reliable and multimodal large language models (LLMs) in French, we collected various social scientific datasets to evaluate models across various tasks. To this effect, we created **BenCSSmark**, a benchmark composed of tasks *from actual computational social science projects*. Bringing together scholars from computer science and the social sciences, BenCSSmark pursues two complementary objectives. First, it aims to encourage NLP scholars to discover the richness of available datasets in the social sciences. Second, it seeks to guide model development by focusing on tasks pertaining to domains that are understudied. Our hope is indeed that this initiative will improve the quality of future models for social scientific tasks, and overall. It is also that BenCSSmark's example will highlight the relevance of social science tasks and lead to a better integration into model evaluations.

2. Benchmarks and Social Sciences: A Limited Interaction

2.1. Related Works

The encounter between AI and the social sciences has proven both rich and productive. On the one hand, social scientists have devoted considerable attention to AI as a cultural and social phenomenon in its own right. They have examined its diffusion, the forms of resistance it generates, and the social, economic, and institutional transformations it brings about (Tubaro et al., 2020; Acemoglu, 2021; Bryan, 2026). They have also examined how AI systems — trained on vast corpora of human-generated data — reproduce, reinterpret, and at times amplify the social norms, biases, and imaginaries embedded in the societies from which these data originate (Garg et al., 2018; Bender et al., 2021).

In addition to these social sciences of AI, researchers have also explored the use of AI in the social sciences, following the distinction proposed by Xu et al. (2024). This includes earlier techniques derived from machine learning (Athey and Imbens, 2019), as well as more recent advances in generative AI (Bail, 2024) and, in particular, large language models (Ziems et al., 2024). Across the social sciences, these tools have inspired new research practices, sparked methodological debates, and fueled growing controversies regarding their epistemological implications and limits (Binz et al., 2025; Boelaert et al., 2025b).

However, despite these fruitful intersections, the development of benchmarks either derived from or tailored to the social sciences has lagged. It is telling that a recent and comprehensive benchmark survey does not include many tasks from these disciplines, even in the section devoted to the social sciences (Ni et al., 2025).

This initial observation does not imply that the social sciences are altogether absent from the natural language processing (NLP) literature on evaluation. Recently, two projects have proposed such benchmarks in adjacent directions. One such initiative is HSSBench (Kang et al., 2025), an extensive

benchmark designed to test multimodal models on tasks related to human knowledge. It presents several thousand questions in six different languages, requiring a model to combine advanced factual knowledge with visual recognition. However, its primary orientation is toward evaluating visual or multimodal models, not models aimed specifically at social science questions.

Another important initiative, in part closer to our own, is the work initiated by Li et al. (2024). The authors aggregated 480 datasets designed to measure "social intelligence." Their tasks are designed to assess a model's capacity to interpret cognitive, situational, and behavioral cues in interactional contexts. While some of these tasks (such as blame attribution, irony or sarcasm detection, nuanced abuse classification, and toxicity moderation) could indeed be relevant to social science research, Li et al.'s benchmark differs in emphasis: it primarily targets dialogue-based understanding, often to improve conversational agents rather than the broader tasks of empirical social inquiry.

Though seemingly more distant, other projects also merit mention. A recent review, for instance, compiled studies that account for the diversity of viewpoints expressed on a given issue and assembled datasets in which multiple, and sometimes conflicting, annotations are preserved (Frenda et al., 2025). This perspectivist approach, which challenges the notion of a unique ground truth or gold standard, draws heavily on insights from the social sciences — and in particular on the idea that evaluative judgments are themselves situated and shaped by experience. As Röttger et al. (2022) have argued, annotation can follow different epistemological logics: some approaches aim to resolve disagreement and converge toward consensus, while others deliberately retain divergence to reflect the plurality of perspectives embedded in language.

2.2. Making Sense of the Absence

Why are social science benchmarks so scarce when it comes to LLM evaluation? Several factors help explain this. First, unlike in machine learning, these disciplines have not historically relied on benchmarks as tools for the cumulative improvement of knowledge. This explains the paradox whereby recent articles systematically report performance metrics or quality assessment, yet no centralized compilation of these evaluations can be found. Assessments are often carried out within individual studies — sometimes using large, publicly available datasets (Ziems et al., 2024) or *ad hoc* collections created for a specific project (e.g., Gilardi et al. 2023). But these datasets are rarely standardized, shared or made available for systematic testing. The absence of consistent data formats and shared repositories thus makes the creation of reproducible benchmarks particularly labor-intensive.

Another reason lies in the concentration of current NLP research on social issues over a limited set of tasks, those that most closely resemble traditional NLP applications. Sentiment analysis, hate speech or toxicity detection, misinformation detection, and stance detection are among the best-developed (Thapa et al., 2025), in large part because they have clear commercial applications that attract funding. By contrast, few social science tasks can be directly framed as commercially relevant. Yet existing benchmarks only represent a fraction of the methodological and analytical needs of the social sciences, whether in text classification, information extraction, or corpus-based analysis. And even within these relatively standardized tasks, persistent difficulties remain. Contextual variation — linguistic, cultural, or situational — often challenges model generalization. For instance, Nogara et al. (2025) demonstrate that Perspective API — often considered state-of-the-art in online hate detection — systematically assigns higher toxicity scores to German-language content than to comparable English inputs, highlighting the uneven cross-linguistic and contextual robustness of contemporary AI systems.

A third difficulty lies in the fact that social science tasks are rarely standardized. On the contrary, research in these disciplines often relies on context-specific categories and tailor-made analytical frameworks. This is partly due to the recurrent debates about the definition of key phenomena (e.g. "social class" or "populism") and disagreements on how best to measure such concepts, which often lead to different operationalizations. More importantly, this is also because social science research questions often require bespoke tools. For instance, in a study of interactions between elected officials and citizens, Claesson (2025) sought to measure the proportion of messages received by politicians that were not hateful or toxic, but also simply "critical", and even "supportive". As no existing tool could accurately capture this distinction, she developed her own classifier. While this poses clear challenges from a knowledge cumulation perspective, it also represents a condition for the inventiveness and interpretive nuance that define these disciplines.

2.3. Social Science Tasks

The preceding discussion points to a broader lack of conceptual clarity regarding what constitutes a social science task — an ambiguity that may have further obscured their absence from widely used benchmarks. In our view, social science tasks are those that exhibit construct validity with respect to

social scientific concepts or questions. This definition implies not only that the task has been employed within the social sciences, but also that it has been regarded as a first-best operationalization of a given approach. In most cases, such tasks have required researchers to produce *ad hoc* expert annotations. However, these annotations are rarely reused or incorporated into existing benchmarks.

It could be argued that most of these tasks would still fall into already existing well-known NLP tasks categories, representing only marginal variations. We contend that their value lies precisely in introducing these variations and nuances.

3. BenCSSmark

3.1. The Initiative

BenCSSmark was created as part of Pantagruel, a collective research project aiming to develop the next generation of large language models (LLMs) in French. The interdisciplinary team includes researchers from the social sciences, who were not brought into the project to focus on the social implications of AI or to audit bias, but rather to integrate tasks directly derived from their own disciplines. Their contribution centers on designing and curating a set of tasks specific to research in the social sciences. The underlying motivation is that as these disciplines increasingly engage with AI tools, they generate datasets that can improve the linguistic and analytical quality of language models.

3.2. Data Collection Strategies

We pursued two strategies simultaneously: the first consisted of producing entirely new data. Common in NLP, this approach involves assembling a team of annotators, training them for a specific task, and generating a dataset. In our case, all annotations were preserved — in addition to the adjudicated decision — in order to reflect the diversity of viewpoints that can emerge around a given task.

The second method, more typical of the social sciences, relies on datasets annotated in great detail by a single individual, often an expert in the relevant domain (Do et al., 2024). Such data cannot be analyzed using conventional procedures, such as inter-annotator agreement metrics. Nevertheless, these annotations are numerous. They are also annotated by experts, and they span a wide range of topics, areas and period. As a result, they constitute an invaluable source for constructing benchmarks that can help develop models better attuned to the interpretive and contextual needs of social scientific research.

3.3. Three Principles for a Benchmark

BenCSSmark was created with three guiding principles in mind aimed to maximize its relevance. The first principle is to better represent the diversity of viewpoints and approaches of social science disciplines. To do so, we draw tasks from underrepresented fields such as sociology and political science. More importantly, for more classical NLP problems such as text classification tasks such as topic, frame or stance detection, we made sure to respect how these disciplines have been approaching and discussing these tasks. To be sure, a social scientist would hardly consider frame detection on social media posts today to be the same task as frame detection on newspaper articles from the first half of the 20th century — at least progress on the former would not be considered relevant for applications on the latter. We also included more unusual tasks originating from these disciplines such as concept detection (searching for instances of populism, of gender, etc.).

The second principle guiding our benchmark is to reflect contextual, socio-cultural and temporal variations. Such variations lie at the heart of social scientific inquiry, which seeks to uncover both enduring regularities and significant shifts in how societies classify situations and name phenomena. Our effort, therefore, consists in assembling data from different data genres and historical periods in order to cover as wide a range of topics, populations, and contexts as possible.

The final principle guiding our work is to preserve, whenever possible, the diversity of annotators' perspectives. For all tasks that we annotated ourselves, we therefore produced two versions of the data: one with a single ground truth (as is commonly done) and one where all single annotations have been kept. These data presents two key advantages. First, it allows us to evaluate task difficulty in a conventional manner — low agreement among annotators may indicate greater ambiguity or complexity, while high agreement suggests clearer boundaries and greater reliability. Second, it contributes to the perspectivist call for producing disaggregated data[1].

For these reasons, our datasets are enriched with relevant metadata. This includes information about the task and disciplinary context, as well as detailed information on the text data and the time period covered. Whenever feasible, we also collect metadata about annotators, enabling their annotations to be situated within their social and contextual backgrounds — a practice that a small number of datasets in the perspectivist literature have begun to adopt (Frenda et al., 2025, p. 1713).

[1] https://pdai.info/

3.4. Data

A crucial aspect of our initiative is the systematic categorization of datasets across multiple facets. Each task is first assigned to a broad task category. The aim here is to characterize the literature related to the task in NLP. Tasks are then also categorized according to the concepts the researchers have been looking to capture. This concept categorization is novel and is introduced to make explicit the specificity of each of these *social science* tasks. Together these two taxonomies look to further exchanges between computer and social scientists. Additional information focuses on the dataset characteristics (data genre, descriptive statistics, temporal scope, modality). This multidimensional organization allows for more precise selection and comparison across contexts.

At the time of writing, BenCSSmark contains 27 datasets, all in French (see Table 1). This number is expected to grow as the project advances and we increase the range of social scientists we reach out to. Beyond presenting our current work, this article aims to invite researchers in the social sciences and natural language processing to contribute to a similar initiative.

Most of the collected tasks fall into common and well-studied task categories such as topic classification, frame detection, hate-speech detection, bias detection, quote detection, or coreference resolution. However, they often deviate from the typical use cases of these categories in terms of the concepts studied and/or the data to which they are applied. Examples include topic classification of music-related articles in the arts & culture sections of the national press; topic-specific frame detection with politically-loaded angles[2]; distinguishing hateful from critical comments in political tweets; detecting biased statements on Wikipedia; detecting unattributed quotes in the press or using a coreference resolution task to measure how much newspapers cite sources with different political leaning.

The benchmark also includes task categories that remain comparatively less formalized, such as concept detection or argumentative strategy detection. For the former the tasks we collected focus on detecting gender-related or social class-related research papers in a corpus of social science paper abstracts. For the latter, the tasks correspond to detecting when opinion pieces on the radio or in politicians' discourses push their arguments using certain types of persuasion strategies.

The remaining task category ("other detection") groups together *ad hoc* tasks which did not fit in any of the common task categories. The concepts these tasks study illustrate how varied and creative social science tasks can be. A notable consequence of the idiosyncrasy of these tasks is that model performance on them is unknown – thus providing novel tests to assess the generalization capacities of models. Examples include detecting when newswriting on music-related topics is prescriptive (e.g. album reviews, tour announcement, interviews), detecting when politicians make reform pledges in their electoral manifestos, or detecting when opinion pieces on the radio make political prophecies.

Regarding the tasks' types, BenCSSmark features binary and multiclass classification, multilabel classification, span detection and coreference resolution. As the benchmark expands we hope to include other types of tasks that may be relevant to the social sciences, such as semantic text similarity or clustering. In doing so we also hope to popularize these other approaches among social scientists.

The text genres cover social media, news, political discourses, broadcast discussions, broadcast news, electoral manifestos and Wikipedia and social science articles. In terms of time period, the datasets currently span the second half of the 20th century onwards. Earlier datasets pose distinctive challenges — archaic vocabulary, shifting social categories, and changes in discourse conventions — that make them particularly valuable for evaluating the robustness and adaptability of modern language models. As for the units of annotation, BenCSSmark contains examples of both sentence, paragraph and full text annotations. The gender concept detection tasks (tasks number 6 and 7) notably produced annotations at both the sentence and the paragraph-level, allowing model performance on each to be compared.

The project currently covers two modalities: written text and speech transcription[3]. Speech transcription poses distinct challenges, notably residual transcription errors — despite substantial recent progress in automatic speech recognition — as well as the complexity of multi-speaker dialogues, which necessitate dedicated formatting and preprocessing procedures.

3.5. The Social Sciences as a Stress Test for Artificial Intelligence

When building a dataset for the social sciences, we are not only intent on improving models performances for social scientific applications. Our aim is also to leverage the richness of their tasks to hopefully advance research on large language models more broadly.

Large language models are often evaluated on tasks that prioritize precision, factual recall, or nar-

[2]For instance, discussing taxation-related topics presenting it as an undue pressure on the wealthy

[3]It is the case for the radio and TV data.

#	Task Category	Type	Concept	Data Genre	Description	Period	Size	Unit	Annotators XP	RA	P
1	Argumentative strategy detection	binary classif.	Appeal to authority	radio	opinion pieces	2017-2023	900	parag.		4	✓
2			Appeal to majority				900	parag.		4	✓
3			Appeal to majority	speech	politicians' speeches	1974-2022	900	parag.		4	✓
4	Bias detection	binary classif.	Non-neutral statement	wikipedia pages	politicians wikis	2002-2024	130	varied	5		✓
5	Bias labeling	3-label classif.	Non-neutral statement bias type	wikipedia pages	politicians wikis	2002-2024	130	varied	1		
6	Concept detection	binary classif.	Gender	academic articles	social science abstracts	2001-2025	2,889	parag.	1		
7			Gender				4,100	sentence	2		
8			Social class				1,990	parag.	2		
9	Coreference resolution	pairwise coref.	Press political source intensity	press	national dailies	1998-2020	200	full text	2	5	
10	Frame detection	3-class classif.	Immigration framing	press	national dailies	2000-2019	1,601	parag.	1		
11			LGBT rights framing				2,996	parag.	1		
12			Taxation framing				2,144	parag.	1		
13	Hate speech detection	binary classif.	Abusive comment	tweets	politics-related tweets	2022-2023	2,152	full text	1		
14			Critical comment				651	full text	1		
15	Other detection	binary classif.	Inclusive language	academic articles	social science abstracts	2001-2025	1,511	parag.	1		
16			Reform pledge	electoral manifestos	French party programs	2015-2024	28,215	sentence	1		
17			Political forecasting	radio	opinion pieces	2017-2023	1,650	parag.		4	✓
18			Music newswriting (prescription detection)	press	national dailies	1998-2023	1,450	full text	1		
19			Supportive comment	tweets	politics-related tweets	2022-2023	1,174	full text	1		
20	Quote detection	span detect.	Press political source diversity	press	national dailies	1998-2020	120	full text	2	5	
21		binary classif.	Unattributed quote	press	national dailies	1945-2018	10,633	sentence	2	3	
22		span detect.	Unattributed quote				10,633	sentence	2	3	
23	Topic classif.	binary classif.	Music-related content	press	national dailies	1998-2023	1,450	full text	1		
24		13-class classif.	News categories	press	national dailies	1945-2022	2,000	full text		1	
25		24-class classif.	Policy issues	tweets	politics-related tweets	2008-2023	6,386	full text	1		
26		3-class classif.	Political newswriting (horserace detection)	press	national dailies	1945-2018	3,843	sentence	2	3	
27	Topic labeling	115-label classif.	Thematic categorization	radio/tv	radio/tv transcripts	1982-2025	2,500	parag.	28		

Table 1: Collected Datasets. **Category** corresponds to common NLP terminology for tasks. **Concept** focuses on what social scientists have been trying to capture. **XP** stands for Experts, **RA** for Research Assistants and **P** stands for Perspectivism and indicates whether individual annotations are available. All tasks come from ongoing or published social science projects. See Table 2 in the appendix for tasks' authors.

row reasoning abilities. Yet these metrics fail to capture one of the most demanding dimensions of human intelligence and diversity: the capacity to interpret meaning in context, to navigate ambiguity, and to arbitrate between conflicting perspectives. The perspectivist literature has started to point out such limits but translating these insights into evaluation metrics remains an open question (one exception being Gordon et al., 2021).

Social science tasks offer precisely this kind of challenge. By confronting models with historically and culturally situated data, multiple view-

points, and conceptually fluid categories, they expose weaknesses that remain invisible in conventional benchmarks. They require models to handle pragmatic nuance, shifting semantics, and perspectival disagreement — dimensions that are central to human communication but peripheral to most NLP evaluations, particularly when it comes to annotation practices.

From this standpoint, BenCSSmark can be viewed as a stress test for AI. It evaluates not only linguistic competence but also interpretive robustness and socio-cognitive flexibility. Performance on social science tasks thus becomes an indicator of a model's broader generalization capacity — its ability to deal with variability, inconsistency, and moral ambiguity inherent in human data. Additionally BenCSSmark also serves as a call for developing novel evaluation metrics that better account for subjectivity or ambiguity in model scoring.

4. Limitations and Future Work

4.1. A First Step Only

This initiative naturally comes with limitations. The first concerns the current scale of the data, which remains limited in both quantity and scope. It is restricted to a limited range of media and forms of expression. Another limitation of BenCSSmark lies in the current nature of the data. At the moment, due to the goals of the project, the datasets are exclusively textual, and only in French. The media covered remain largely confined to conventional formats (such as newspapers and official speeches) and well-studied domains (such as social media posts).

Yet this initiative should first be read as an invitation: an open call for social scientists to engage with and contribute to benchmarks, in the spirit of a collaborative project designed to serve the collective advancement of both NLP and the social sciences. Our aim with this project is thus not to build a definitive benchmark, but rather to draw the attention of both NLP practitioners and social science researchers to the importance of more sustained exchange between the two communities.

NLP researchers could benefit from training, selecting, and validating their models on the numerous and fine-grained data that already exist within the social sciences, which often simply lack visibility. Because they pertain to fundamentally human activities involving language, of central interest to linguists (and to models), and because they introduce diversity and nuance into computational tasks, social science data constitute a valuable reservoir for NLP. Conversely, for social scientists, having access to models trained on tasks aligned with their research needs would clearly be advantageous. By identifying the specific needs of these disciplines — which constitute a substantial share of contemporary research — we can help ensure that they are adequately represented and integrated into the development of future models and benchmarks.

Another limitation is that BenCSSmark remains closed, at least for the time being. Our initial objective was to provide a fully open repository. However, we were compelled to revise this plan, as much of the data is either proprietary (e.g., newspaper archives) or may contain personally identifiable information. In both cases, hosting the data within a shared infrastructure would have exposed us to legal and ethical liabilities, and we therefore decided against this option for the time being.

Moreover, as large language model (LLM) developers increasingly incorporate existing datasets into their training corpora, benchmarks risk becoming rapidly outdated — a challenge that currently affects many widely used evaluation frameworks. These issues are not specific to our project, and addressing them will require the development of innovative infrastructures and governance models in the future.

4.2. Avoiding the Risks of Benchmarkisation

Another risk inherent in this initiative is that it may reproduce the very shortcomings commonly associated with benchmarking — particularly the distortions that arise from the excessive reliance on standardized evaluation metrics. These limitations have been extensively documented in the literature.

Benchmarks, by construction, tend to narrow the definition of what counts as relevant to what can be measured. They may foster optimization toward the metric rather than toward the phenomenon of interest (Goodhart's Law), leading models to overfit to test sets instead of improving in generalizable ways. Moreover, when benchmarks become central instruments of evaluation, they risk disciplining research agendas, privileging technical improvement over theoretical or conceptual innovation (Hooker, 1995; Paullada et al., 2021).

Benchmarks can thus become unproductive. This fact has been extensively discussed in the NLP community (Liao et al., 2021; Weidinger et al., 2025; Raji et al., 2021). It raises acute questions for a set of disciplines as diverse and interpretively rich as the social sciences. Excessive standardization could undermine the pluralism of approaches that characterizes the discipline, silencing context, uncertainty, and disagreement — precisely the elements that constitute its epistemic strength.

A straightforward way to guard against this drift is to design benchmarks that incorporate multiple labels — such as time period, discipline, cultural

area, and task type — thereby enabling evaluations to be filtered according to relevant criteria rather than collapsed into a single aggregate score. Another protection is to continually expand such benchmarks with new datasets, thereby introducing new research questions and perspectives over time.

5. Conclusion

Benchmarks have long guided progress in large language model (LLM) research. They focus community efforts on standardized tasks and enable direct comparisons between models. However, tasks that are not included in benchmarks are often neglected: models are not optimized for them and their performance on such tasks remains largely unknown. This has notably been the case for *social science tasks*, despite the wide range of datasets produced by social scientists in their research.

As an initial step toward addressing this gap, we introduced BenCSSmark, a benchmark composed of social science tasks from ongoing social science research projects or published papers. Our objectives are threefold: to help conceptualize social science tasks and highlight their diversity; to provide an initial benchmark tailored to them; and to foster dialogue between computer scientists and social scientists by engaging with their distinct research traditions.

Expanding LLM benchmarking to better represent the social sciences presents opportunities for both communities. For social scientists, it offers a way to shape model development toward tools that better address their research questions. For computer scientists, it opens access to a growing body of carefully annotated datasets. By making the social sciences count, our hope is that we will see improvements in both fields.

6. Acknowledgements

This research has been partially funded by the French National Research Agency (ANR project "PANTAGRUEL", ANR-23-IAS1-0001). It is also partly supported by Hi! PARIS and the ANR/France 2030 program (ANR-23-IACL-0005). It also received government funding managed by ANR under France 2030, reference ANR-23-IACL-0006.

7. Bibliographical References

Daron Acemoglu. 2021. Harms of AI. Technical report, National Bureau of Economic Research.

Susan Athey and Guido W Imbens. 2019. Machine learning methods that economists should know about. *Annual Review of Economics*, 11(1):685–725.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. 2021. Program Synthesis with Large Language Models. ArXiv:2108.07732 [cs].

Christopher A. Bail. 2024. Can Generative AI improve social science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Marcel Binz, Stephan Alaniz, Adina Roskies, Balazs Aczel, Carl T. Bergstrom, Colin Allen, Daniel Schad, Dirk Wulff, Jevin D. West, Qiong Zhang, Richard M. Shiffrin, Samuel J. Gershman, Vencislav Popov, Emily M. Bender, Marco Marelli, Matthew M. Botvinick, Zeynep Akata, and Eric Schulz. 2025. How should the advancement of large language models affect the practice of science? *Proceedings of the National Academy of Sciences*, 122(5):e2401227121.

Julien Boelaert, Samuel Coavoux, Estelle Delaine, Altaïr Despres, Sibylle Gollac, Narguesse Keyhani, Adèle Momméja, and Étienne Ollion. 2025a. La part du genre : Genre et approche intersectionnelle dans les revues de sciences sociales françaises au xxie siècle. *Actes de la recherche en sciences sociales*, 258-259(3-4):126–145.

Julien Boelaert, Samuel Coavoux, Étienne Ollion, Ivaylo Petev, and Patrick Präg. 2025b. Machine bias. how do generative language models answer opinion polls?1. *Sociological Methods & Research*, 54(3):1156–1196.

Kevin A. Bryan. 2026. The economic impacts of artificial intelligence: A multidisciplinary, multi-book review. *Journal of Economic Literature*, 64(1):281–300.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, et al. 2021. Evaluating Large Language Models Trained on Code. ArXiv:2107.03374 [cs].

Annina Claesson. 2025. Le prix de la visibilité : Une analyse computationnelle des interactions en

ligne avec des parlementaires. *Revue Française de Science Politique*, 75(3):549–579.

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. ArXiv:2110.14168 [cs].

Kate Crawford and Trevor Paglen. 2021. Excavating AI: the politics of images in machine learning training sets. *AI & Society*, 36:1105–1116.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. 2021. The Benchmark Lottery. ArXiv:2107.07002 [cs].

Salomé Do, Étienne Ollion, and Rubing Shen. 2024. The augmented social scientist: Using sequential transfer learning to annotate millions of texts with human-level accuracy. *Sociological Methods & Research*, 53(3):1167–1200.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2025. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*, 59(2):1719–1746.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. 2021. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning AI with shared human values. In *International Conference on Learning Representations*. Poster.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*. Poster.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

John N Hooker. 1995. Testing heuristics: We have it all wrong. *Journal of heuristics*, 1(1):33–42.

Hugging Face. 2025. Open llm leaderboard. https://huggingface.co/spaces/open-llm-leaderboard. Accessed: 2025-10-15.

Florian Jaton. 2021. *The Constitution of Algorithms: Ground-Truthing, Programming, Formulating*. The MIT Press.

Zhaolu Kang, Junhao Gong, Jiaxu Yan, Wanke Xia, Yian Wang, Ziwen Wang, Huaxuan Ding, Zhuo Cheng, Wenhao Cao, Zhiyuan Feng, Siqi He, Shannan Yan, Junzhe Chen, Xiaomin He, Chaoya Jiang, Wei Ye, Kaidong Yu, and Xuelong Li. 2025. HSSBench: Benchmarking Humanities and Social Sciences Ability for Multimodal Large Language Models. ArXiv:2506.03922 [cs].

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2024. Analyzing dataset annotation quality management in the wild. *Computational Linguistics*, 50(3):817–866.

Bernard Koch, Emily Denton, Alex Hanna, and Jacob G Foster. 2021. Reduced, reused and recycled: The life of a dataset in machine learning research. ArXiv:2112.01716 [cs.LG].

Minzhi Li, Weiyan Shi, Caleb Ziems, and Diyi Yang. 2024. Social Intelligence Data Infrastructure: Structuring the Present and Navigating the Future. ArXiv:2403.14659 [cs].

Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. 2021. Are we learning yet? a meta review of evaluation failures across machine learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Yang Liu, Jiahuan Cao, Chongyu Liu, Kai Ding, and Lianwen Jin. 2024. Datasets for large language models: A comprehensive survey. ArXiv:2402.18041 [cs.CL].

Dieuwertje Luitse, Tobias Blanke, and Thomas Poell. 2025. AI competitions as infrastructures of power in medical imaging. *Information, Communication & Society*, 28(10):1735–1756.

Stephen Mayhew, Terra Blevins, Shuheng Liu, Marek Šuppa, Hila Gonen, Joseph Marvin Imperial, Börje F. Karlsson, Peiqin Lin, Nikola Ljubešić, LJ Miranda, Barbara Plank, Arij Riabi, and Yuval Pinter. 2024. Universal NER: A gold-standard multilingual named entity recognition benchmark. 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 4322–4337, Mexico City, Mexico. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Shiwen Ni, Guhong Chen, Shuaimin Li, Xuanang Chen, Siyi Li, Bingli Wang, Qiyao Wang, Xingjian Wang, Yifan Zhang, Liyang Fan, Chengming Li, Ruifeng Xu, Le Sun, and Min Yang. 2025. A Survey on Large Language Model Benchmarks. ArXiv:2508.15361 [cs].

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. ArXiv:2207.04672 [cs.CL].

Gianluca Nogara, Francesco Pierri, Stefano Cresci, Luca Luceri, Petter Törnberg, and Silvia Giordano. 2025. Toxic Bias: Perspective API Misreads German as More Toxic. *Proceedings of the International AAAI Conference on Web and Social Media*, 19(1):1346–1357.

Etienne Ollion, Rubing Shen, Ana Macanovic, and Arnault Chatelain. 2023. Chatgpt for text annotation? mind the hype. SocArXiv preprint.

Will Orr and Edward B. Kang. 2024. AI as a Sport: On the Competitive Epistemologies of Benchmarking. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1875–1884, Rio de Janeiro, Brazil. Association for Computing Machinery.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336.

Inioluwa Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the Everything in the Whole Wide World Benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Rubing Shen. 2024. *The politics of newswriting : three essays on how journalists cover politics*. Ph.D. thesis, Institut d'études politiques, Paris. Thèse de doctorat en sociologie dirigée par Cointet, Jean-Philippe et Ollion, Étienne.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, et al. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. ArXiv:2206.04615 [cs].

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. ArXiv:2210.09261 [cs].

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Surabhi Adhikari, Hariram Veeramani, Mehwish Nasim, and Usman Naseem. 2025. Large language models (LLM) in computational social science: prospects, current

state, and challenges. *Social Network Analysis and Mining*, 15:article number 4.

Paola Tubaro, Antonio A Casilli, and Marion Coville. 2020. The trainer, the verifier, the imitator: Three ways in which human platform workers support artificial intelligence. *Big Data & Society*, 7(1).

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: a stickier benchmark for general-purpose language understanding systems. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 3266–3280, Red Hook, NY, United States. Curran Associates Inc.

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, Brussels, Belgium. Association for Computational Linguistics.

Laura Weidinger, Inioluwa Deborah Raji, Hanna Wallach, Margaret Mitchell, Angelina Wang, Olawale Salaudeen, Rishi Bommasani, Deep Ganguli, Sanmi Koyejo, and William Isaac. 2025. Toward an evaluation science for generative ai systems. ArXiv:2503.05336 [cs.AI].

Ruoxi Xu, Yingfei Sun, Mengjie Ren, Shiguang Guo, Ruotong Pan, Hongyu Lin, Le Sun, and Xianpei Han. 2024. AI for social science and social science of AI: A survey. *Information Processing & Management*, 61(3):103665.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024. WorldValuesBench: A large-scale benchmark dataset for multi-cultural value awareness of language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17696–17706, Torino, Italia. ELRA and ICCL.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can Large Language Models Transform Computational Social Science? *Computational Linguistics*, 50(1):237–291.

A. Additional Table

#	Category	Type	Concept	Data	Contact	Status	Reference
1	Argumentative strategy detection	binary classif.	Appeal to authority	radio	Yacine Chitour	ongoing work	
2			Appeal to majority		Yacine Chitour	ongoing work	
3			Appeal to majority	speech	Yacine Chitour	ongoing work	
4	Bias detection	binary classif.	Non-neutral statement	wikipedia pages	Victor Planche	ongoing work	
5	Bias labeling	3-label classif.	Non-neutral statement bias type	wikipedia pages	Victor Planche	ongoing work	
6	Concept detection	binary classif.	Gender	academic articles (parag.)	Julien Boelaert	published	Boelaert et al. (2025a)
7			Gender	academic articles (sentence)	Julien Boelaert	published	Boelaert et al. (2025a)
8			Social class		Julien Boelaert	published	Boelaert et al. (2025a)
9	Coreference resolution	pairwise coref.	Press political source intensity	press	Emma Bonutti	ongoing work	
10	Frame detection	3-class classif.	Immigration framing	press	Rubing Shen	PhD chapter	Shen (2024, chap. 3)
11			LGBT rights framing		Rubing Shen	PhD chapter	Shen (2024, chap. 3)
12			Taxation framing		Rubing Shen	PhD chapter	Shen (2024, chap. 3)
13	Hate speech detection	binary classif.	Abusive comment	tweets	Annina Claesson	published	Claesson (2025)
14			Critical comment		Annina Claesson	published	Claesson (2025)
15	Other detection	binary classif.	Inclusive language	academic articles	Julien Boelaert	published	Boelaert et al. (2025a)
16			Institutional reform pledge	electoral manifestos	Frédéric Gonthier	ongoing work	
17			Music newswriting (prescription detection)	press	Samuel Coavoux	ongoing work	
18			Political forecasting	radio	Yacine Chitour	ongoing work	
19			Supportive comment	tweets	Annina Claesson	published	Claesson (2025)
20	Quote detection	span detect.	Press political source diversity	press	Emma Bonutti	ongoing work	
21		binary classif.	Unattributed quote	press	Rubing Shen	PhD chapter	Shen (2024, chap. 2)
22		span detect.	Unattributed quote	press	Rubing Shen	PhD chapter	Shen (2024, chap. 2)
23	Topic classification	binary classif.	Music-related content	press	Samuel Coavoux	ongoing work	
24		13-class classif.	News categories	press	Francesco Colonna	ongoing work	
25		24-class classif.	Policy issues	tweets	Malo Jan	ongoing work	
26		3-class classif.	Political newswriting (horserace detection)	press	Salomé Do	published	Do et al. (2024)
27	Topic labeling	115-label classif.	Thematic categorization	radio/tv	Nicolas Hervé	INA internal	

Table 2: Authors and References for the Collected Datasets. **#** uniquely identifies each dataset. **Category** corresponds to common NLP terminology for tasks. **Concept** focuses on what social scientists have been trying to capture. **Data** presents data genre and the unit of text annotation in parenthesis when needed for disambiguation. **Contact** provides a link to one of the data author's websites. **Status** is the research project status.