International Journal of Social Trends
Volume 4, Issue 5 (Sept-Oct; 2026)

Review Article · Sociology and Social Research Methods

From Topic Models to Large Language Models: A Methodological Review of Computational Text Analysis in Sociology

Mikkelsen Weber [1], **Oliver Noah** [2]

[1] University of Bucharest, ROMANIA
[2] Eötvös Loránd University, HUNGARY

ABSTRACT

Sociologists have long treated text as evidence of culture, institutions and social relations. Over the past two decades, computational methods have made it possible to analyse corpora of a size that no team of human coders could read, and the arrival of large language models (LLMs) has changed the field again. This review traces that development through three methodological eras (bag-of-words models, word embeddings and generative models), explains what each family of methods measures, and assesses the evidence on how well LLMs perform the tasks sociologists care about. Zero-shot annotation with instruction-tuned models has matched or exceeded crowd workers on several classification tasks at a fraction of the cost, with intercoder agreement above 90 percent in one widely cited study, but accuracy varies widely across tasks and datasets and falls on newer, less familiar material. Simulating survey respondents with LLMs reproduces some aggregate patterns while misrepresenting the opinions of particular groups. We argue that the main contribution of LLMs is to change the economics of measurement rather than to replace interpretive judgement, and we propose a validation workflow built on explicit constructs, gold-standard samples, versioned prompts, bias-corrected downstream estimation and full reporting. The review closes with a discussion of reproducibility, data protection and the use of open models in sociological research.

Keywords: Computational Sociology; Text as Data; Topic Models; Word Embeddings; Large Language Models; Measurement Validity; Content Analysis

INTRODUCTION

Text has always been one of sociology's richest sources of evidence. Newspapers, court records, letters, interviews, organisational reports and, more recently, social-media posts record how people and institutions describe the world, and in doing so they reveal the categories, frames and meanings through which social life is organised. For most of the discipline's history, analysing such material meant close reading or manual content analysis, in which trained coders apply a codebook to a sample of documents [1]. These methods remain the foundation of interpretive rigour, but they scale poorly. A coder can read hundreds of documents; a digitised archive or a social-media platform can contain billions.

Computational text analysis emerged to close this gap. Early applications in political science treated words as data to estimate the positions of parties [2, 3], and a landmark review summarised both the promise and the pitfalls of automated content analysis, warning that all quantitative models of language are wrong but some are useful, and that validation is indispensable [4]. In sociology, topic models were taken up as tools

for studying culture and meaning [5, 6], and the broader programme of computational social science framed digital text as part of a new observational infrastructure for studying society [7]. Reviews in economics and sociology have since mapped how text can be used for measurement, prediction and discovery [8, 9], and an influential textbook reframed the field around research tasks rather than algorithms [10].

The arrival of large language models, culminating in the public release of ChatGPT in November 2022, has changed the field faster than any previous development. Models that can follow natural-language instructions now classify, summarise, translate and generate text with little or no task-specific training data [11, 12]. Early studies suggested that they could replace crowd workers for many annotation tasks [13], simulate the opinions of demographic groups [14], and support new forms of qualitative and quantitative inquiry [15, 16]. Others warned of opacity, bias, environmental cost and threats to reproducibility [17, 18].

This review is written for sociologists who want to use these tools well. In it, we (i) place LLMs within the longer history of computational text analysis, (ii) explain what each family of methods actually measures, (iii) assess the evidence on LLM performance for sociological tasks, and (iv) propose a validation workflow and reporting standards that preserve the discipline's commitment to valid measurement. Throughout, we treat computational methods as instruments whose properties must be understood before their outputs can be interpreted.

Box 1. Four tasks for text as data

Following recent methodological work [10, 19], it helps to separate four research tasks. Discovery uses text to find new concepts or categories, as in exploratory topic modelling or computational grounded theory [20]. Measurement assigns documents to known categories or positions on a scale. Prediction uses text to forecast an outcome. Causal inference uses text as a treatment, an outcome or a confounder in a causal design.

Each task asks different things of a method. A model that is excellent for discovery may be poor for measurement, and a highly accurate predictor may be useless for causal inference if its errors are correlated with the treatment. Choosing a method begins with choosing the task.

THREE ERAS OF COMPUTATIONAL TEXT ANALYSIS

Figure 1 summarises the development of methods discussed in this review. The periods are not strict: older methods remain in use and are often the right choice. They do, however, mark successive changes in how text is represented, from counts of words, to positions in a geometric space, to the internal states of models trained to predict and generate language.

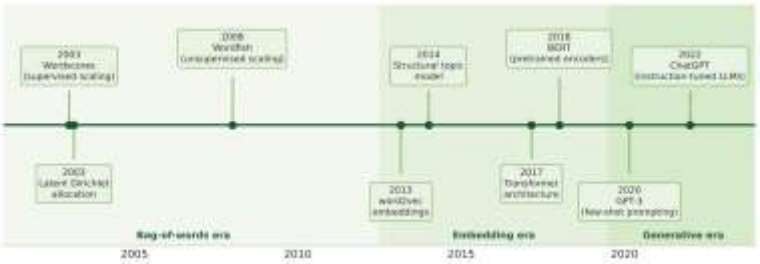

Figure 1: Milestones in computational text analysis relevant to the social sciences, grouped into three methodological eras. Source: compiled by the authors from the publications cited in this section.

The Bag-of-Words Era

The first era represented each document as a vector of word counts and ignored word order. Within this representation, two broad strategies developed. Supervised scaling methods such as Wordscores used reference texts with known positions to place other texts on the same scale [2], while unsupervised scaling methods such as Wordfish estimated positions from the relative frequency of words across documents without reference texts [3]. For discovering themes, latent Dirichlet allocation (LDA) modelled each document as a mixture of topics and each topic as a distribution over words [21].

Sociologists found topic models especially congenial. A study of newspaper coverage of U.S. government arts funding argued that topic models share key assumptions with sociological theories of culture: meaning is relational, the same words take on different meanings in different contexts, and culture is organised into frames that can be identified from patterns of co-occurrence [5]. The structural topic model extended LDA by allowing document-level covariates, such as the author's party or the date of publication, to shape topic prevalence and content, which made it possible to test hypotheses about how discourse varies across groups and over time [22]. Applications ranged from open-ended survey responses [22] to the study of censorship, where an analysis of millions of Chinese social-media posts showed that the government permitted criticism of officials but removed posts that could facilitate collective action [23].

The bag-of-words era also produced important methodological cautions. Results from unsupervised models can change substantially with seemingly minor preprocessing choices, such as stemming, removing stop words or dropping rare terms, and researchers rarely report how sensitive their findings are to these decisions [24]. The interpretation of topics depends on human reading of the most probable words and documents, which introduces a subjective step that must be documented and, ideally, validated [4].

Scale and the Digitised Archive

Alongside these modelling developments, the digitisation of archives changed what counts as a corpus. An analysis of millions of digitised books introduced the term “culturomics” for the quantitative study of cultural trends through word frequencies over centuries, tracking phenomena such as the adoption of new technologies, the fading of collective memory and the effects of censorship [25]. The study demonstrated the scale now available, but it also prompted debate about the composition of such corpora: digitised book collections over-represent certain genres, publishers and periods, so changes in word frequencies can reflect

changes in what was collected rather than changes in culture. For sociologists, this debate is a reminder that sampling questions do not disappear with scale. They move from the selection of documents by the researcher to the selection of documents by libraries, platforms and digitisation projects.

The Embedding Era

The second era represented words as dense vectors learned from their contexts, so that words used in similar contexts receive similar vectors. Word2vec and GloVe made it practical to train such embeddings on large corpora [26, 27]. Because relations between words are encoded as directions in the vector space, embeddings opened new ways of measuring cultural associations. A study of the meanings of social class projected words onto cultural dimensions such as affluence, education and cultivation, defined by differences between antonym pairs, and traced how these dimensions changed across the twentieth century using embeddings trained on books [28]. Embeddings trained on a century of text quantified changing gender and ethnic stereotypes and correlated those changes with occupational and demographic data [29], and an influential study showed that standard embeddings reproduce human-like biases measured by implicit association tests [30].

Embeddings also have measurement pitfalls. Results depend on the corpus, the training algorithm and its settings, and the stability of nearest-neighbour relations varies across random initialisations. A systematic evaluation for applied research found that pretrained embeddings often perform as well as locally trained ones for many tasks, but that the choice should be validated against human judgements of similarity for the specific corpus and question [31]. Sentence-level embeddings from transformer encoders later made it possible to represent whole documents in the same way [32], and topic-modelling approaches built on such embeddings combined clustering of document vectors with class-based term weighting [33].

The Generative Era

The third era rests on the transformer architecture, which uses attention mechanisms to represent each word in the context of all the others [34]. Pretrained encoders such as BERT learned contextual representations from large corpora and could be fine-tuned for specific classification tasks with relatively small labelled datasets [35]. Generative models trained to predict the next word showed that, at sufficient scale, they could perform many tasks from instructions and a few examples supplied in the prompt, without any fine-tuning [11]. Further training on human feedback produced instruction-following models that respond to plain-language requests [12]. For social scientists, the practical change is that a researcher can now describe a coding scheme in words and obtain annotations from a model, rather than training a classifier on thousands of hand-coded examples.

Table 1: Families of computational text methods and what they measure

Method family	Representation	Typical sociological use	Key validation requirement
Dictionaries	Counts of pre-specified words	Sentiment, moral language, specific themes	Check that word lists capture the construct in the corpus

Method family	Representation	Typical sociological use	Key validation requirement
Scaling models	Word frequencies with reference texts or latent positions	Ideological or policy positions	Face and convergent validity against expert placements
Topic models	Documents as mixtures of word distributions	Discovering frames and themes; change over time	Semantic coherence; human interpretation; robustness to preprocessing
Word embeddings	Dense vectors learned from context	Cultural associations, stereotypes, meaning change	Stability across runs; agreement with human similarity judgements
Fine-tuned transformers	Contextual representations adapted to a task	High-accuracy classification with labelled data	Held-out accuracy by subgroup and period
Prompted LLMs	Instructions and examples in natural language	Zero- or few-shot annotation, summarisation, extraction	Agreement with expert codes; prompt sensitivity; version control

WHAT LARGE LANGUAGE MODELS CAN DO FOR SOCIOLOGY

Evaluations of LLMs for social-science tasks have accumulated rapidly. We group them under three headings: annotation and measurement, simulation of human respondents, and support for qualitative and exploratory analysis.

Annotation and Measurement

The most influential early evaluation compared ChatGPT with crowd workers on Amazon Mechanical Turk and with trained research assistants across four samples of tweets and news articles totalling 6,183 documents [13]. The tasks included judging whether a text was relevant to content moderation, the stance it took, its topic and the frame it used. Across datasets, ChatGPT's zero-shot accuracy exceeded that of crowd workers by about 25 percentage points on average, and its intercoder agreement, measured by running the model twice, exceeded that of both crowd workers and trained annotators for every task. The cost per annotation was below USD 0.003, about thirty times cheaper than the crowd platform. Figure 2 shows two of the reported results.

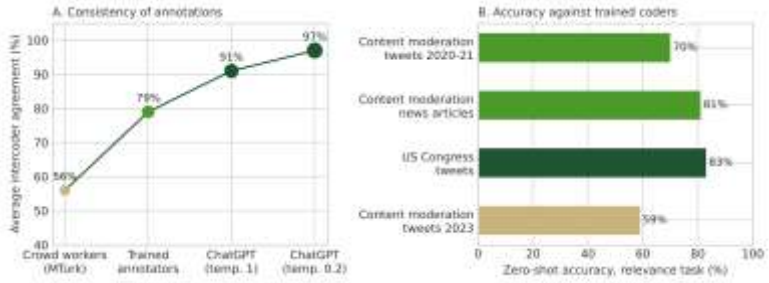

Figure 2: Results reported by Gilardi et al. Panel A shows average intercoder agreement for crowd workers, trained annotators and ChatGPT at two temperature settings. Panel B shows ChatGPT's zero-shot accuracy on the binary relevance task in each of four datasets, using trained annotators as the benchmark. Data source: [13].

Two features of these results deserve emphasis. First, high agreement is not the same as accuracy. The model's agreement with itself was very high, especially at a low temperature setting, but its accuracy against trained coders ranged from 59 percent to 83 percent on the comparatively simple relevance task, and it was lowest on the most recent dataset, which consisted of tweets from 2023 (Figure 2B). Consistency can hide systematic error. Second, the benchmark was the researchers' own trained annotators, so the study shows that the model reproduced expert coding more closely than crowd workers did, not that it captured the construct better than experts. A broad evaluation of LLMs across dozens of computational social-science tasks reached a similar conclusion: models performed well on many classification tasks, especially those with clear definitions, but fell short of fine-tuned models on tasks requiring specialised knowledge or subtle judgements, and the authors recommended using LLMs to augment rather than replace human annotation [16].

A practical implication concerns the choice between prompting and fine-tuning. Prompting requires no labelled training data and can be revised quickly, which makes it attractive for exploratory measurement and for constructs that evolve as a project develops. Fine-tuning a smaller transformer encoder [35] on a few thousand expert labels typically produces a classifier that is cheaper to run on very large corpora, more stable over time because its parameters are fixed, and easier to share and archive. A common and efficient strategy combines the two: an LLM produces initial labels that human coders correct, the corrected labels train a fine-tuned model, and a held-out expert sample evaluates the final classifier. This strategy uses the strengths of each approach while keeping the final measurement instrument transparent and reproducible [10, 16].

Simulating Respondents and “Silicon Samples”

A more ambitious proposal is to use LLMs as stand-ins for survey respondents. By conditioning a model on demographic and political backstories drawn from real surveys, one study generated “silicon samples” whose responses reproduced fine-grained patterns in voting behaviour and the associations among political attitudes found in human data [14]. The authors called this property algorithmic fidelity and suggested that such samples could help researchers pilot questions and explore hypotheses. Others have argued that LLMs might replace human participants for some psychological studies [36].

The evidence counsels caution. An analysis of the opinions expressed by language models found substantial misalignment with the views of many U.S. demographic groups, and showed that instruction tuning shifted models toward the opinions of particular groups, such as more liberal and more educated respondents [37]. Simulated respondents tend to show less variance than real ones, and their answers depend on prompt wording in ways that are hard to predict. For sociology, a discipline concerned precisely with the distribution of views across groups, these limitations are serious. Silicon samples may be useful for designing instruments or generating hypotheses, but they are not a substitute for data collected from people.

Support for Qualitative and Exploratory Work

LLMs also offer help with tasks that sit closer to the interpretive tradition, such as summarising interviews, proposing candidate codes, identifying passages relevant to a theme or translating material. Computational grounded theory proposed a three-step cycle in which computational pattern detection, guided deep reading and pattern confirmation alternate [20]; LLMs can accelerate each step, for example by surfacing passages for close reading or by applying a refined codebook to a full corpus. Earlier work combining text analysis with network methods showed how computational tools can connect what organisations say to how audiences respond [38]. A broad assessment of generative AI in social science identified opportunities in survey design, simulation of social interactions and analysis of unstructured data, alongside risks of bias, homogenisation of research and over-reliance on proprietary tools [15].

MEASUREMENT VALIDITY IN THE AGE OF LLMS

The central methodological question for sociologists is not whether LLMs can produce labels, but whether those labels validly measure the constructs in a theory. Figure 3 places the main families of methods by two properties that bear on this question: how much labelled data the researcher must supply, and how sensitive the representation is to context.

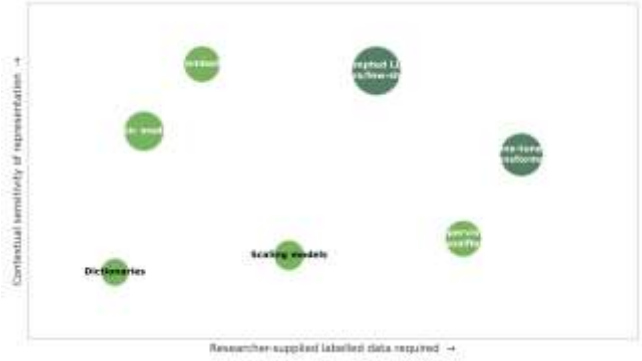

Figure 3: Conceptual positioning of text-analysis methods by the amount of researcher-labelled data they require and the contextual sensitivity of their representations. Bubble size indicates breadth of current use in social-science research. Positions are illustrative. Source: authors' design.

Construct Validity

A prompt is a measurement instrument, and it inherits all the problems of instrument design. A request to label a post as “populist” or “uncivil” will be interpreted through the model's training data, which reflects

how such terms are used online rather than how a particular theory defines them. Content analysis has long required explicit codebooks, coder training and reliability testing [1], and these requirements apply with equal force when the coder is a model. A codebook written for humans is a good starting point for a prompt, and disagreements between the model and human coders often reveal ambiguities in the codebook itself.

Reliability, Sensitivity and Drift

Reliability has new dimensions. Outputs vary with the sampling temperature, the wording and order of instructions, the examples included in the prompt and, for proprietary services, the version of the model behind an interface that can change without notice. Gilardi and colleagues found that a lower temperature increased consistency without reducing accuracy [13], which suggests a simple default, but prompt sensitivity must be tested for each application. Model drift poses a distinctive threat to reproducibility: a study using a commercial model through an application programming interface may be impossible to replicate once that model is retired. This is a central reason for the argument that open models, whose weights can be archived and rerun, are the ethical choice for social science [66].

Bias in Measurement

Language models absorb associations present in their training data, including stereotypes about gender, race and other categories [17, 29, 30]. When researchers use such models as measurement instruments, these associations can become measurement error correlated with the very groups sociologists study. For example, a classifier of "toxic" language may flag dialects associated with particular communities at higher rates. Validation must therefore report performance separately for relevant subgroups, not only in aggregate. The same embedding methods that reveal stereotypes in text [42, 65] can be used to audit models for them.

From Labels to Estimates

Even accurate labels can produce biased estimates when they are used in downstream analysis. If a model's errors are correlated with a covariate of interest, regression coefficients estimated on model labels will be biased, and standard errors will understate uncertainty. Design-based supervised learning addresses this by combining model predictions for all documents with expert labels for a randomly selected subset, producing estimates with valid confidence intervals even when the model is imperfect [39]. The approach formalises an intuition that runs through the validation literature: a small, carefully drawn gold-standard sample is not just a check on the model but an input to the estimation itself.

Table 2: Threats to validity when using LLMs for sociological measurement, with mitigations

Threat	How it arises	Mitigation
Construct mismatch	Model applies everyday meanings of theoretical terms	Explicit codebook in the prompt; pilot against expert codes
Prompt sensitivity	Small wording changes alter labels	Test multiple prompts; report variation; fix final prompt before analysis

Threat	How it arises	Mitigation
Model drift	Proprietary model versions change or are retired	Use open or version-pinned models; archive prompts and outputs
Subgroup bias	Training-data associations correlate errors with groups	Report accuracy by subgroup; audit with known-bias tests
Downstream bias	Correlated errors bias regression estimates	Bias-corrected estimation using a random expert-labelled subset
Temporal shift	Accuracy falls on material newer or different from training data	Validate on samples from each period and source
Data protection	Sending sensitive text to external services	Local models; de-identification; ethics review of data flows

A VALIDATION WORKFLOW

Drawing these considerations together, we propose a six-step workflow (Figure 4). It is designed to be proportionate: a quick exploratory analysis need not complete every step, but any study that reports quantitative findings based on model-generated labels should.

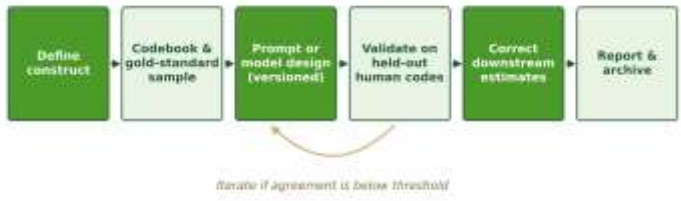

Figure 4: A validation workflow for LLM-assisted measurement in sociological research. The loop indicates iteration on the prompt or model when agreement with the gold standard falls below a pre-specified threshold. Source: authors' design.

Step 1: Define the construct. State the concept to be measured, its theoretical definition and the unit of analysis. Decide whether the task is discovery, measurement, prediction or causal inference [10].

Step 2: Build a codebook and gold-standard sample. Write a codebook as if for human coders, draw a random sample of documents, and have at least two trained coders label it, reporting their agreement [1]. Reserve part of the sample as a held-out test set.

Step 3: Design and version the prompt or model. Translate the codebook into instructions, choose between zero-shot prompting, few-shot prompting and fine-tuning, fix the model version and settings, and record everything needed to rerun the analysis. Where possible, prefer open models that can be archived [18].

Step 4: Validate on held-out human codes. Compare model labels with the held-out gold standard overall, by subgroup and by period. Set an acceptance threshold in advance. If performance falls short, revise the prompt or the codebook and return to Step 3, keeping the test set untouched until the final evaluation.

Step 5: Correct downstream estimates. Use the expert-labelled sample to correct for classification error when estimating quantities of interest, rather than treating model labels as ground truth [39].

Step 6: Report and archive. Report the prompt, model, settings, validation results and their uncertainty, and archive inputs and outputs, subject to data-protection constraints. Table 3 lists the minimum reporting items.

Table 3: Minimum reporting checklist for studies that use LLMs to measure social constructs

Item	What to report
Construct and task	Theoretical definition; unit of analysis; whether the task is discovery, measurement, prediction, or causal inference
Model	Name, version or checkpoint, access route, date of use, and whether weights are open
Prompt	Full text of instructions and examples; temperature and other settings; number of runs
Gold standard	Sampling procedure; number of documents; coder training; human intercoder agreement
Validation	Accuracy, precision, recall or agreement against held-out codes, overall and by subgroup and period
Downstream analysis	How classification error was accounted for in estimates and confidence intervals
Sensitivity	Results under alternative prompts or models
Data governance	Where data were processed; de-identification; ethics approval

ETHICAL AND INSTITUTIONAL CONSIDERATIONS

Data Protection and Consent

Sociological text data often contain personal and sensitive information: interview transcripts, case files, messages, medical narratives. Sending such material to a commercial service may breach ethics approvals or data-protection law, and the terms under which providers retain or use inputs vary. Local deployment of open models avoids many of these problems, and its costs have fallen. Where external services are used, de-identification and data-processing agreements should be in place before analysis begins.

Concentration of Power and Reproducibility

If social research comes to depend on a small number of proprietary models, the discipline cedes control over its instruments to companies whose priorities differ from those of science. Critics have also

emphasised the environmental and financial costs of training very large models and the difficulty of documenting their training data [17]. Open models with documented training procedures, shared evaluation datasets and institutional computing resources are the most direct responses [18]. Journals and funders can encourage them by requiring the reporting items in Table 3.

Interpretive Authority

A final consideration is disciplinary. Sociology has a long tradition of interpretive analysis that values context, reflexivity and attention to the meanings people themselves give to their words. There is a risk that cheap automated labelling will crowd out close reading and fieldwork, or that the categories built into models will quietly replace those developed through sociological theory. The workflow proposed here is meant to guard against this by keeping human judgement at the points that matter most: defining constructs, writing codebooks, reading the cases where model and coders disagree, and interpreting results.

Training and Research Capacity

The adoption of LLMs also has implications for how sociologists are trained. Earlier waves of computational methods required programming skills and statistical training that were unevenly distributed across departments, which concentrated computational work in well-resourced institutions. Instruction-following models lower some barriers, because a researcher can obtain annotations by describing a coding scheme in plain language. They raise others, because evaluating model outputs, correcting downstream estimates [39] and deploying open models locally still require technical expertise. Graduate curricula that combine measurement theory, content-analysis methodology [1] and practical computing will be better placed to use these tools responsibly than curricula that treat them as a shortcut. Shared institutional infrastructure, including secure computing environments and curated evaluation datasets, can reduce inequalities in access between institutions and between researchers in different countries.

APPLICATIONS ACROSS SOCIOLOGICAL FIELDS

To show how these methods map onto substantive research, Table 4 summarises representative applications discussed in this review and the methodological lessons each offers.

Table 4: Representative applications of computational text analysis in sociology and neighbouring disciplines

Field	Study focus	Method	Methodological lesson
Sociology of culture	Newspaper framing of public arts funding	Topic model	Topic models align with relational theories of meaning
Social stratification	Meanings of class across the twentieth century	Word embeddings	Cultural dimensions can be measured as directions in vector space
Gender and ethnicity	Century-long change in stereotypes	Word embeddings	Text-based measures can be validated against census data

Field	Study focus	Method	Methodological lesson
Political sociology	Censorship of social media in China	Automated text analysis at scale	Large-scale measurement can test theories of state control
Public opinion	Open-ended survey responses	Structural topic model	Covariates can be built into topic estimation
Social movements and organisations	How advocacy messages generate conversation	NLP combined with network analysis	Text and network data are complementary
Survey methodology	Simulating respondents with LLMs	Conditioned generative models	Aggregate fidelity can coexist with group-level misrepresentation

Note. Studies listed in row order: [5]; [28]; [29]; [23]; [22]; [38]; [14].

Two patterns emerge from these applications. First, the most influential studies paired computational measurement with an explicit theoretical question and an independent check on validity, such as comparing embedding-based measures with census occupation data [29] or reading samples of removed posts [23]. Second, the methods were chosen for the task rather than for novelty. A topic model was the right tool for exploring frames [5], whereas measuring change in specific associations called for embeddings [28]. LLMs extend the toolkit but do not change this logic.

PRACTICAL GUIDANCE

Choosing a Method in Practice

Researchers often ask which method to use. The answer depends on the task, the size and type of corpus, the availability of labelled data and the stakes of error. For exploratory work on a corpus whose themes are not known in advance, topic models and embedding-based clustering remain efficient and transparent, provided that results are checked for robustness to preprocessing [24] and interpreted through close reading of representative documents. For measuring a well-defined construct across a large corpus, prompted LLMs now offer a fast route to a first set of labels, but those labels should be validated and, where accuracy matters, used to train or check a fine-tuned classifier whose behaviour is fixed and documented [16]. For tracing changes in associations or meanings over long periods, embeddings trained on period-specific corpora remain the most direct instrument [28, 29]. For hypotheses about how discourse differs across groups, models that incorporate covariates directly, such as the structural topic model, avoid a two-step procedure in which measurement error is ignored [22].

Two further considerations cut across these choices. The first is cost, which now includes not only researcher time but also computing resources and, for commercial models, per-token fees that scale with corpus size. The second is the audience for the findings. Results intended to support claims about specific groups or to inform policy require stronger validation, fuller reporting and more conservative uncertainty estimates than exploratory analyses intended to generate hypotheses. Matching the rigour of validation to

the stakes of the conclusions is a principle that predates computational methods [1] and remains the best guide to their use.

CONCLUSION

Large language models represent the latest and most dramatic step in a long development of computational text analysis. They make it possible to apply detailed coding schemes to very large corpora at low cost, and early evaluations show that, for many well-defined tasks, they reproduce expert coding more closely than crowd workers do [13, 16]. They also raise familiar problems in new forms: construct validity, reliability, bias, reproducibility and the protection of research participants. The history reviewed here suggests that each new representation of text brought both genuine gains and new pitfalls, and that the field advanced when methodological enthusiasm was matched by validation [4, 24, 31]. For sociologists, the appropriate stance is neither wholesale adoption nor rejection. LLMs are best understood as measurement instruments that change the economics of research: they make the expensive part of content analysis cheap, but they do not make the theoretical part optional. Treated in this way, with clear constructs, gold-standard validation, bias-corrected estimation and transparent reporting, they can greatly extend the range of questions that sociology can answer with text.

REFERENECS

[1] Potharaju, B. (2026). Infrastructure-level intelligence: Embedding AI into data movement and validation layers. *International Journal of Computational and Experimental Science and Engineering*.

[2] Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review, 97*(2), 311–331.

[3] Karim, N. B., Tolosana, D., & Simko, R. (2025). Responsible and Legally Compliant Enterprise Artificial Intelligence in Small-Business Lending: Governance, Explainability, and Fair-Lending Safeguards for Early-Intervention Recommendation Engines. *International Journal of Modern Computing, 8*(1), 55-66.

[4] Slapin, J. B., & Proksch, S.-O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science, 52*(3), 705–722.

[5] Potharaju, A. (2026). Human-guided artificial intelligence for adaptive compliance automation in financial institutions. *Journal of Computational Analysis and Applications*, 35(4), 57–68.

[6] Mallempati, A., Jaladi, D. S., & Jegajothi, B. (2026). HDLF-DGC: A hybrid deep learning framework for intelligent cybersecurity and data governance in distributed web systems. In *2026 Third International Conference on Innovations in Cybersecurity and Data Science (ICICDS)* (pp. 16–21). IEEE.

[7] Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*(3), 267–297.

[8] Aluri, G. N. C., & Mupparapu, H. K. (2021). Spring Batch framework patterns for large-scale data processing in cloud-native microservices: A GCP deployment study. *International Journal of AI, BigData, Computational and Management Studies*, 2(4), 130–142.

[9] DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics, 41*(6), 570–606.

[10] Mohr, J. W., & Bogdanov, P. (2013). Introduction—Topic models: What they are and why they matter. *Poetics, 41*(6), 545–569.

[11] Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Van Alstyne, M. (2009). Computational social science. *Science, 323*(5915), 721–723.

[12] Karim, N. B., & Kovac, T. (2025). Cloud-Based Big-Data Architectures for AI-Driven Financial Planning and Analysis and Early Risk Signaling in Resource-Constrained Community Financial Institutions. *The Metascience*, *3*(3), 29-40.

[13] Movva, C. K., & Mupparapu, H. K. (2026). MVVM architectural patterns for enterprise Android kiosk applications: State management and lifecycle handling. *International Journal of Emerging Trends in Computer Science and Information Technology*, 7(2), 320–330.

[14] Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature, 57*(3), 535–574.

[15] Potharaju, B., & Ravipudi, J. (2023). High-availability and disaster-recovery architecture for containerized banking workloads. *International Journal of Data Engineering and Intelligent Computing*, 6(1), 1–32.

[16] Rahman, M. S., Bhowmik, P. K., Hossain, B., Tannier, N. R., Amjad, M. H. H., Chouksey, A., & Hossain, M. (2023). Enhancing fraud detection systems in the USA: A machine learning approachto identifying anomalous transactions. *Journal of Economics, Finance and Accounting Studies*, *5*(5), 145-160.

[17] Potharaju, A. (2026). Policy-driven authorization architectures for modern financial regulatory platforms. *Journal of International Crisis and Risk Communication Research*, 9(4), 36–44.

[18] Evans, J. A., & Aceves, P. (2016). Machine translation: Mining text for social theory. *Annual Review of Sociology, 42*, 21–50.

[19] Movva, C. K., & Manohar, V. (2021). OCF protocol implementation for smart home IoT applications: An Android SDK development case study. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 2(1), 89–96.

[20] Shammah, R. S., Sultana, S., Faruq, O., Kowser, S., Uddin, A., & Haque, S. A. (2026, July). Data-Driven Decision Support Framework for Engineering and IT Project Management Using Integrated Business Analytics and Management Information Systems. In *2026 IEEE International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)* (pp. 1-6). IEEE.

[21] Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

[22] Yuvaraj, N., Kumar, M. S., & Manoj, M. S. (2026). Unified agentic AI architecture for data lineage, governance, and predictive workflow automation in enterprise systems. In *2026 IEEE 18th International Conference on Computational Intelligence and Communication Networks (CICN)* (pp. 1817–1828). IEEE.

[23] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33* (pp. 1877–1901). Curran Associates.

[24] Aluri, G. N. C., & Manohar, V. (2026). Model context protocol servers as orchestration layers for generative AI agents in enterprise software delivery. *International Journal of AI, BigData, Computational and Management Studies*, 7(2), 337–345.

[25] Kumar, M. S., Yuvaraj, N., & Manoj, M. S. (2026). Intelligent orchestration of master data and privacy compliance via agentic AI for customer lifecycle optimization. In *2026 IEEE 18th International Conference on Computational Intelligence and Communication Networks (CICN)* (pp. 1715–1723). IEEE.

[26] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35* (pp. 27730–27744). Curran Associates.

[27] Mallempati, A., & Jaladi, D. S. (2023). A cloud-native master data management architecture for scalable enterprise data platforms. *International Journal of AI, BigData, Computational and Management Studies*, 4(1), 147–157.

[28] Potharaju, B., & Sajja, R. R. (2024). Modernizing US federal regulatory reporting infrastructure for scalability, resilience, and audit readiness. *International Journal of Data Engineering and Intelligent Computing*, 7(1), 1–36.

[29] Rahman, M. S., Kabir, A. A., Reon, M. A. H., & Rashid, S. U. (2025). From Financial Transactions to Clinical Records: A Unified Machine Learning Methodology for Real-Time Anomaly Detection and Institutional Decision Support In High-Stakes Domains. *International Journal of Information Technology and Management Information Systems (IJITMIS)*, *16*(6), 23-42.

[30] Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, *120*(30), e2305016120.

[31] Karim, N. B., Kovac, T., & Song-Chun, Z. (2025). Model Risk Management for Adaptive Credit Early-Warning Systems in Supervised SmallBusiness Lending: Aligning Machine-Learning Recalibration

[32] with Federal Model-Risk Guidance and the NIST AI Risk Management Framework. *International Journal of Acta Informatica*, *4*(1), 80-94.

[33] Potharaju, A. (2021). Designing dynamic questionnaire engines for risk-based account review workflows. *International Journal of Applied Data Science & Modern Computing*, 4(1), 1–37.

[34] Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337–351.

[35] Bail, C. A. (2024). Can generative AI improve social science? *Proceedings of the National Academy of Sciences, 121*(21), e2314021121.

[36] Yuvaraj, N., & Kumar, M. S. (2025). Agentic AI for zero-touch customer experience: A governance-constrained framework for autonomous service systems. *American International Journal of Computer Science and Technology*, 7(2), 100–111.

[37] Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics*, *50*(1), 237–291.

[38] Rashid, S. U., Al Muzakkiruddin, S., & Møller, C. (2026). Detecting and Correcting Temporal Drift in Credit-Risk and Fraud-Detection Models at Community Banks, Credit Unions, and CDFIs: An Automated Model-Risk-Monitoring Framework Below the Supervisory Threshold. *The Computertech*, 26-37.

[39] Movva, C. K., & Aluri, G. N. C. (2025). Enterprise mobile payment gateway architecture on IoT handheld devices: Security and reliability design. *International Journal of AI, BigData, Computational and Management Studies*, 6(1), 232–241.

[40] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (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* (pp. 610–623). ACM.

[41] Kumar, M. S., & Yuvaraj, N. (2024). Predictive customer experience orchestration using governed data pipelines and intelligent service signals. *International Journal of Emerging Trends in Computer Science and Information Technology*, 5(1), 206–215.

[42] Spirling, A. (2023). Why open-source generative AI models are an ethical choice for social science. *Nature*, *616*(7957), 413.

[43] Potharaju, A. (2026). Embedded compliance architecture for real-time financial systems modernization. *Journal of Information Systems Engineering and Management*, 11(2s), 1147–1158.

[44] Zuberi, N., Rahman, M. M., Rashid, S. U., Ahsan, M. N., Rahman, M. D., & Mollah, M. H. O. R. (2026). A robust machine learning framework for detecting temporal drift in financial fraud prevention. *Scientific Reports*.

[45] Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021). Machine learning for social science: An agnostic approach. *Annual Review of Political Science*, *24*, 395–419.

[46] Shahiduzzaman, M., Sourav, M. S. A., Kabtia, M., Shammah, R. S., Islam, M. A., Chakraborty, P., ... & Adnan, M. (2026). Attention-guided quantum-inspired extreme learning machine with grey wolf optimization for rare-attack detection in IIoT networks. *Discover Artificial Intelligence*.

[47] Nelson, L. K. (2020). Computational grounded theory: A methodological framework. *Sociological Methods & Research*, *49*(1), 3–42.

[48] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

[49] Siddiqui, M. I. H., Rahman, M. S., Kabir, A. A., Mahmud, F. U., Rashid, S. U., & Shammah, R. S. (2023). Comparative analysis of explainable machine learning models for cancer classification using cytological features. *Journal of Medical and Health Studies*, *4*(5), 110-150.

[50] Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science, 58*(4), 1064–1082.

[51] Aluri, G. N. C., & Manohar, V. (2025). Generative AI integration patterns for enterprise workflow automation: A practitioner framework. *International Journal of Emerging Research in Engineering and Technology*, 6(3), 165–172.

[52] Shammah, R. S., Rahman, I., Guria, Z. M., Kowser, S., Faruq, O., & Morshed, N. (2025, December). Scalable Machine Learning Architecture for Securing IoT Networks with Digital Governance in the AI Era of Cyber-Resilient Smart Cities. In *2025 International Conference on AI-Driven STEM Education and Learning Technologies (AISTEMEDU)* (pp. 1-7). IEEE.

[53] Manohar, V., & Mupparapu, H. K. (2025). Scalable data pipeline architecture for real-time supply chain analytics using PySpark and Snowflake. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 6(3), 203–212.

[54] King, G., Pan, J., & Roberts, M. E. (2013). How censorship in China allows government criticism but silences collective expression. *American Political Science Review, 107*(2), 326–343.

[55] Mallempati, A. (2024). Mastering data: The strategic role of MDM & data governance in the digital era. *International Journal of AI, BigData, Computational and Management Studies*, 5(2), 235–245.

[56] Potharaju, A. (2026). Microservices-based modernization of financial regulatory systems. *International Journal of Computer Science and Engineering Innovations*, 2(1), 68–85.

[57] Krippendorff, K. (2018). *Content analysis: An introduction to its methodology* (4th ed.). SAGE Publications.

[58] Rahman, M. M., Computer, M. M. R., Das, N., Uddin, M. S., Shammah, R. S., & Rahman, M. M. (2025, May). An Investigation of Face Detection and Feature Extraction Techniques in Deepfake Detection Using a Multi-Model Machine Learning Approach. In *2025 3rd International Conference on Business Analytics for Technology and Security (ICBATS)* (pp. 1-8). IEEE.

[59] Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis, 26*(2), 168–189.

[60] Mupparapu, H. K., & Movva, C. K. (2020). Serverless migration patterns for compliance-critical financial applications on Microsoft Azure. *International Journal of Emerging Research in Engineering and Technology*, 1(2), 85–96.

[61] Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331*(6014), 176–182.

[62] Potharaju, B., & Ravipudi, J. (2025). Automated data validation and resilient infrastructure design for enterprise regulatory reporting platforms. *International Journal of Computer Science and Engineering Innovations*, 1(1), 46–64.

[63] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26* (pp. 3111–3119). Curran Associates.

[64] Jaladi, D. S., Mallempati, A., & Jegajothi, B. (2026). A hybrid federated privacy-aware deep learning framework for secure data governance in cloud systems. In *2026 4th International Conference on Sustainable Computing and Smart Systems (ICSCSS)* (pp. 1629–1636). IEEE.

[65] Karim, N. B., & Moreau, L. (2026). Quantifying the Prospective Impact of AI-Assisted Early Intervention in US Small-Business Lending: Default Avoidance, Credit-Access Preservation, and Small-Business Employment Effects Across Community Lending Portfolios. *The Metascience*, *4*(2), 33-44.

[66] Potharaju, B. (2026). Architecting national-scale regulatory reporting platforms: A high-availability infrastructure framework for compliance-critical systems. *Journal of International Crisis & Risk Communication*.