Humanities & Social Sciences
Communications

ARTICLE

https://doi.org/10.1057/s41599-025-04503-w

OPEN

Machine-assisted quantitizing designs: augmenting humanities and social sciences with artificial intelligence

Andres Karjus [1,2] ✉

The increasing capacities of large language models (LLMs) have been shown to present an unprecedented opportunity to scale up data analytics in the humanities and social sciences, by automating complex qualitative tasks otherwise typically carried out by human researchers. While numerous benchmarking studies have assessed the analytic prowess of LLMs, there is less focus on operationalizing this capacity for inference and hypothesis testing. Addressing this challenge, a systematic framework is argued for here, building on mixed methods quantitizing and converting design principles, and feature analysis from linguistics, to transparently integrate human expertise and machine scalability. Replicability and statistical robustness are discussed, including how to incorporate machine annotator error rates in subsequent inference. The approach is discussed and demonstrated in over a dozen LLM-assisted case studies, covering nine diverse languages, multiple disciplines and tasks, including analysis of themes, stances, ideas, and genre compositions; linguistic and semantic annotation, interviews, text mining and event cause inference in noisy historical data, literary social network construction, metadata imputation, and multimodal visual cultural analytics. Using hypothesis-driven topic classification instead of "distant reading" is discussed. The replications among the experiments also illustrate how tasks previously requiring protracted team effort or complex computational pipelines can now be accomplished by an LLM-assisted scholar in a fraction of the time. Importantly, the approach is not intended to replace, but to augment and scale researcher expertise and analytic practices. With these opportunities in sight, qualitative skills and the ability to pose insightful questions have arguably never been more critical.

---

[1] School of Humanities, Tallinn University, Tallinn, Estonia. [2] Estonian Business School, Tallinn, Estonia. ✉email: andres.karjus@tlu.ee

Introduction

Developments in generative large language models (LLMs, sometimes dubbed AI) have broadened their applicability to various research tasks. Of particular interest to the humanities and social sciences (H&SS) is the capacity to use them as on-demand classifiers and inference engines. Classifying texts or images for various properties has been available via supervised machine learning (ML) for a while. Yet, the necessity to train (or tune pretrained) models for every single variable of interest on sufficiently large sets of labeled examples may have been one factor hampering the wider adoption of ML tools as research instruments (in humanities but also text-involved industries, especially in low-resource scenarios). Unsupervised ML (word embeddings, latent topic modeling) has allowed for explorative approaches but often requires complex text preprocessing and convoluted pipelines to use for confirmatory inference, with their latent spaces opaque to interpret.

In-context (zero-shot) learning, now widely available via generative instructable LLMs, does not require the laborious labeling of training data. Their instruction-driven usage naturally lends itself to distilling researcher expertise into scalable big data annotators. Many current LLMs, when properly instructed or tuned, can perform on par with human analysts, and occasionally even better than non-expert assistants (De Paoli 2023; Gilardi et al. 2023; Pilny et al. 2024; Rathje et al. 2024; Törnberg 2023; Webb et al. 2023; Ziems et al. 2023). There are, however, several challenges to such “collaborative intelligence” (Brinkmann et al. 2023; González-Bailón et al. 2023; Mollick 2024; Schleiger et al. 2023). One is validation and quality control: an LLM with a high public benchmark score cannot be blindly expected to perform in specific research tasks and contexts. There is operationalization and transparency: variables generated via classification usually constitute quantitative data, necessitating principled inference pipelines and statistical modeling to ensure the reliability of subsequent interpretations and claims. There is little discussion yet in the recent benchmarking-centered literature on how to best operationalize the potentially very large resulting datasets of extracted variables for hypothesis testing while controlling for machine error rates.

Combining automated classification and statistics is common in some H&SS fields, yet underutilized in others. Qualitative approaches or quantification without systematic testing are both nonviable when it comes to big data. This contribution takes a step beyond the performance testing focus of preceding LLM literature in H&SS and suggests a framework for machine-assisted research, with explicated steps of systematic unitization, qualitative annotation (by humans or machines), and quantitative inference. It is argued to be a more practical alternative compared to past approaches like content analysis. Exemplary case studies are presented to demonstrate its application in tandem with current LLMs as instructable assistants.

Related zero-shot LLM applicability research. This contribution builds upon and complements other recent work in LLM applicability research. This is to denote the exploration of the feasibility and performance of pretrained LLMs as research and analytics tools—as distinct from the ML domain of LLM engineering. ML and NLP-supported research in the (digital) humanities and (computational) social sciences is nothing new. However, until recently, a typical machine-assisted text-focused research scenario would have involved either training a supervised learning classifier (or fine-tuning a pretrained LLM, see, e.g. Rosa et al. 2023; Majumder et al. 2020) on a large set of annotated examples for a given task, or using output vectors from a word or sentence embedding LLM like BERT for clustering or other tasks (e.g. Fonteyn 2021; Sen et al. 2023).

Recent advances in instruction-tuned LLM technology make a difference: it has become feasible to use them for on-demand data classification and annotation, enabling multivariable machine-assisted research designs. In a zero-shot scenario, a generative model is “instructed” to produce output given an input (prompt) including in some form both the analysis instructions and a unit of data (in the few-shot case, also some expected outputs). The generated text can be then parsed and quantified as necessary. This obviates the need for laborious annotation work to create large training or tuning sets for every specific task, question, or variable, which are typically numerous in research dealing with complex linguistic, cultural, or societal topics. Zero-shot does not necessarily outperform fine-tuning or bespoke architectures, even of smaller older models (Ollion et al. 2023; Ziems et al. 2023), but is easier to implement for the aforementioned reason. Tuning or adaptation approaches (Dettmers et al. 2023; Hu et al. 2021) can still be more efficient on larger or repeated tasks.

Another contributing factor to recent LLM popularity, both as chatbots and classifiers, is arguably accessibility. Running very large LLMs requires powerful hardware, while various cloud API services, including that of OpenAI’s GPT (generative pretrained transformer model), have made them accessible, albeit at a cost, to those who do not have access to hardware or skills to operate it. To be clear though: the following discussion pertains to LLMs (e.g. GPT, Llama, Mistral), but not LLM-driven chatbot services (e.g. ChatGPT, Copilot), which are typically not well suited for systematic data analysis.

All of this has attracted attention across various research communities well beyond NLP. Interested parties include the H&SS, which makes sense: LLMs are getting good at text-to-text tasks, and analysis in these fields often involves converting or translating complex text (in the broad semiotic sense) into some form of discrete textual codes, annotations, or taxonomies for further analysis. In a large multidisciplinary benchmarking exercise, Ziems et al. (2023) show how several LLMs perform well across various (English-language) annotation and classification benchmarks. Gilardi et al. (2023) compared the performance of the then-current GPT-3.5 to Amazon Mechanical Turk workers on text classification tasks and found that it outperformed crowdworkers on accuracy and reliability, at a fraction of the cost (see also Huang et al. 2023; Törnberg 2023; Wu et al. 2023a). Others, to name a few, have focused on performance testing on tasks in domains like discourse annotation (De Paoli 2023; Fan and Jiang 2024; Rytting et al. 2023), event data coding (Overos et al. 2024), constructions grammar and frame semantics (Torrent et al. 2023), metalinguistic and reasoning abilities (Beguš et al. 2023; Chi et al. 2024), zero-shot translation (Tanzer et al. 2024), medical and psychology research applications (Demszky et al. 2023; Palaniyappan 2023; Wang et al. 2023a), text retrieval and analytics (Zhu et al. 2023), sentiment, stance, bias and affiliation detection (Nadi et al. 2024; Törnberg, 2023; Wen-Yi et al. 2024; Zhang et al. 2024), including applications to languages beyond just English (Buscemi and Proverbio 2024; Mets et al. 2024; Rathje et al. 2024). Research is artificial cognition has also been explicitly comparing machine and human behavior (Acerbi and Stubbersfield 2023; Binz and Schulz, 2023; Futrell et al. 2019; Taylor and Taylor 2021).

Related feature analytic and mixed methods approaches. This contribution advocates for a general-purpose quantitizing-type research design (QD), consisting of converting each systematically unitized data point (e.g. sentence) into a fixed number of predetermined categorical or numerical variables, followed by statistical modeling of the inferred variables (which is common in some disciplines but not widely so in H&SS). The terminological variability surrounding this approach necessitates a brief review. "Quantitization" or "quantizing" (Fetters et al. 2013; Hesse-Biber 2010; Sandelowski et al. 2009) is a useful term to distinguish the process of annotating data with the explicit purpose of subsequent quantification—from annotation for any other purposes. "Coding" is also frequently used, but unhelpfully also has many other meanings. The qualitative conversion step may range from quick annotation to labor-intensive critical or discourse analysis of every unit of data (Banha et al. 2022; Fofana et al. 2020). Importantly: the research question is answered primarily based on the results of the quantification step (e.g. statistical modeling), and not the data annotation step, nor by simply looking at the coded variables (as is common in some quasi-quantifying approaches).

Combining qualitative and quantitative, it might be useful to situate QD as mixed methods, although the majority of the latter is mixed mostly in the sense of using multiple data types (and a method for each), including sequential, concurrent, convergent, triangulating designs (Hesse-Biber 2010; Huynh et al. 2019; O'Cathain et al. 2010; Tashakkori and Teddlie 2010). Variants similar to QD have also been referred to as "integrated" (Creamer 2018; O'Halloran et al. 2019; Tashakkori and Teddlie 2010), or "integration through data transformation" (not to be confused with "transformative mixed methods"; Mertens 2008), "qualitative/quantitative" (Young and Jaganath 2013), and "converting" (Creamer 2018). Qualitative comparative analysis exemplifies an extreme transformation approach that involves reducing (even potentially very complex) data into binary truth table variables (Kane and Kahwati 2023; Vis 2012). Parks and Peters (2023) discuss a "dialogical" machine-assisted framework which is not quantitizing but also relies on NLP tools. "Distant reading" in digital humanities (Moretti 2013) and "ousiometrics" (Fudolig et al. 2023) are machine-assisted approaches, but typically rely on counting words or topic clusters and not manual coding.

A mixed quantitizing approach where the annotation step is referred to as "coding" is found within content analysis (Banha et al. 2022; Pilny et al. 2024; Schreier 2012), but subsequent quantification of the coded variables is not considered "a defining criterion" of CA (Krippendorff 2019). It also encompasses interpretative-qualitative approaches (Hsieh and Shannon 2005), and quantification limited to counting or limited statistical testing (cf. Morgan 1993; Schreier 2012). Limited quantification is occasionally found in discourse analysis (e.g. O'Halloran et al. 2019). Political event coding also applies coding in this sense (Overos et al. 2024; Schrodt and Van Brackle 2013). Thematic analysis does too but typically without statistical modeling of the resulting distributions (cf. Braun and Clarke 2012; Trahan and Stewart 2013), which is a problem (see "Method details and statistical considerations" section). This includes recent LLM-powered proposals (De Paoli 2023; Hau 2024).

Issues with quantification, aggregation, and biased sampling affect reliability and replicability and lead to spurious results (Parks and Peters 2023). Going forward, approaches using quantification but in a limited manner, either by employing impressionistic claims ("more," "less," "some") without actual quantification, or counting without addressing uncertainty—will be referred to as quasi-quantifying. This contribution emphasizes the need for systematic statistical modeling to estimate uncertainty and deal with confounding variables, interactions, multicollinearity, and repeated measures (more common in the humanities than commonly assumed). Though often dismissed in small-sample humanities research, these issues cannot be ignored when applying automation to operationalize large datasets.

One domain where the approach of combining qualitative coding with subsequent quantification is widespread, if not the default, is usage-based (corpus-based, variationist, cognitive) linguistics, with roots in componential analysis and structural semantics (Goodenough 1956; Nida 1979). It is often not referred to as a specific design, but has been called "usage feature analysis," (sometimes prepended with "multi-factorial"; Glynn and Fischer 2010), or instead "behavioral profiles" (Gries and Divjak 2009). The laborious quantitizing or coding, especially if requiring expert linguistics knowledge, is typically conducted by the researchers themselves. Coding or "classification" schemes may include standard variables like grammatical categories from past literature (Szmrecsanyi et al. 2014) or be developed for a given research question (cf. Glynn 2010). Developing standardized coding schemes or taxonomies can also be the goal itself, as in branches of psychology using similar methods (Hennessy et al. 2016). Unlike some of the disciplines mentioned above, a great deal of attention is paid to rigorous statistical procedures in the quantitative modeling step (Gries 2015; Winter 2020; Wolk et al. 2013).

A machine-assisted quantitizing design (MAQD). The framework discussed in the following is just that: the QD or feature-analytic design, generalized beyond linguistics, with augmentation by capable machines to solve the scaling problem, and emphasizing the need for good practices in unitizing, agreement or error estimation, and principled statistical analysis. It is suitable where the data are qualitative (text, images, etc.) but can be quantitized (annotated, coded, converted) into one or more categorical or numeric variables for quantitative modeling. The suggestion is to think as machines such as LLMs as tools or (narrow) AI assistants that scale expert analysis to larger datasets—but not as "oracles" or "arbiters," in the taxonomy of Messeri and Crockett (2024). The human researcher is encouraged to carry out the interpretation of the quantified end result. A typical pipeline can be summarized as follows, further illustrated in Fig. 1; see the "Method details and statistical considerations" section and the Supplementary Information (SI) for details on these steps.

1. Research question, hypothesis, or explorative goal; data collection.
2. Coding scheme with application instructions; unitization principles.

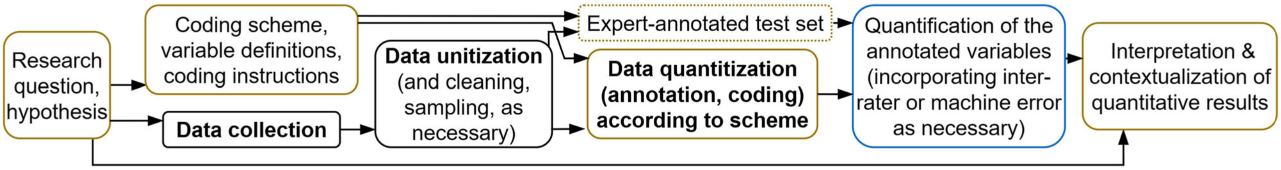
Research question, hypothesis
Coding scheme, variable definitions, coding instructions
Data collection
Data unitization (and cleaning, sampling, as necessary)
Expert-annotated test set
Data quantitization (annotation, coding) according to scheme
Quantification of the annotated variables (incorporating inter-rater or machine error as necessary)
Interpretation & contextualization of quantitative results

Fig. 1 A typical QD pipeline. Qualitative elements are outlined in yellow and quantitative in blue. Steps, where machine assistance (ML, LLMs, or otherwise) may be applied, are in bold, including the quantitization step. Annotating a smaller additional test set is optional but strongly recommended if using either multiple human annotators or a machine.

3. Unitization of data (into meaningful units of analysis); sampling and filtering if necessary.
4. Qualitative annotation (quantitizing) of each unit according to a coding scheme of one or more variables. If delegated to a machine, then also: acquiring or annotating a test set.
5. Quantitative (typically statistical) analysis of the inferred variables, their relationships, and uncertainty. Incorporation of the rates of (human) disagreement or (machine) error in the uncertainty estimation.
6. Qualitative interpretation of the quantitative modeling (potentially in combination with examples from data or theory).

What this contribution does. As discussed above, most of this pipeline is already conveniently standard to many. Applying ML is nothing new in some fields of H&SS, although usually not conceptualized as mixed or quantitizing methods. This is a practical proposal for the wider adoption of one tried and tested design, well suited for maximizing gains from automation. It is hoped to be helpful as a framework that would be easily referenceable, implementable, replicable, and teachable. It can function as an end-to-end analytic framework, or, being theory-agnostic, integrated into existing methodologies such as content or thematic analysis. This contribution has three goals: filling one research gap and contributing to two others. Primarily, it describes and encourages the adoption of the QD, given its advantages over alternatives. The "Method details and statistical considerations" section (and its extension in the SI) discusses the approach in the context of other common designs and describes its components, the importance of statistical modeling and confidence estimation, and how to incorporate machine (or human) error rates in the former.

The general QD can be applied using just human annotators. However, as shown in the case studies below, now readily available automation yields scalability to magnitudes of data that might be unfeasible in purely qualitative or human-annotation paradigms. Size alone should never be the goal in science, but larger data (if properly sampled) can lead to more representative findings and greater statistical power needed to model the complexities of human language, culture, and societies.

Second, the supporting case studies exemplify the application of MAQD principles, while also complementing current LLM applicability research in H&SS which has been mostly focused on large languages like modern English and clean contemporary data. The case studies include nine languages: Estonian, Finnish, German, Italian, Japanese, Latin, Russian, Turkish, and English in four varieties (nineteenth century US, eighteenth century UK, contemporary standard US, and social media US American). Third, the case studies illustrate applying the MAQD and complement related benchmarking literature discipline-wise, with experiments covering several H&SS domains—linguistics, sentiment and discourse analysis, literature and translation, media and film studies, history, social network science, lexicography, and discussing applications of multimodal models in visual cultural analytics. The examples include replications of past research and synthetic examples created for this contribution. Most exemplify the MAQD pipeline, others include practical tasks like LLM-driven data cleaning and content filtering, frequent prerequisites to applying any QD-type approach, especially if the data source is a larger general sample such as a corpus.

Unlike most related research (Rathje et al. 2024; Ziems et al. 2023), the focus here is intentionally not on public benchmarks or shared tasks. One reason is data contamination (Aiyappa et al. 2023), that current LLMs trained likely on open internet data may well include publicly available datasets. In the one NLP benchmark utilized here (one highly unlikely to cause contamination), the zero-shot approach scores 1.5-2x above the state of the art (at the time of writing). Relying on public benchmarks would also be misaligned with the proposed framework, which advocates for smaller but expert-annotated, task-specific test sets—to be used to estimate machine error rates, in turn incorporated in subsequent statistical estimates; but also to compare and choose models (including fine-tuned and personal or research group specific models; see "Discussion" section). The code and test sets are made public, to complement the current benchmarking scene and to provide an easy starting point.

What this contribution is not: three disclaimers. Some disclaimers are in order, as "AI" has attracted a significant uptick of public attention and corporate hype in the last few years. First, we will not be explicitly dealing here with topics like data copyright, related ethical issues, possible model biases, environmental concerns, machine "psychology," "AGI," or AI tractability and theoretical limitations. These issues have been and will be discussed elsewhere (Bender et al. 2021; Binz and Schulz 2023; Rooij et al. 2024; Feng et al. 2023; Hagendorff et al. 2024; Liesenfeld et al. 2023; Lund et al. 2023; Motoki et al. 2023; Novozhilova et al. 2024; Ollion et al. 2024; Tomlinson et al. 2024). There is also a growing literature on what LLMs are not or should not be capable of (Asher et al. 2023; Dentella et al. 2023; Dinh et al. 2023; Miceli-Barone et al. 2023; Sclar et al. 2023). Here the focus is pragmatic: using suitable machines as research tools where applicable, guided by expert evaluation.

Second, this is not about replacing researchers or research assistants (cf. Erscoi et al. 2023), but about augmenting, capacitating, and empowering, while promoting transparent and replicable research practices. Human labor does not scale well, machines do; human time is valuable and machine time is cheap. This is about reducing repetitive labor and increasing time for meaningful work. Ziems et al. (2023) suggest that "LLMs can augment but not entirely replace the traditional [computational social science] research pipeline." Indeed, the largest gains are likely to be made from empowering expert humans with powerful machine annotators and assistants, not assembly line automation of science (Lu et al. 2024). This proposal places qualitative thinking and expert knowledge front and center, requiring designing meaningful hypotheses, coding schemes, and analysis instructions, expert annotation of evaluation datasets, and final contextualization of the results of the statistical step.

Finally, the LLM test results and classification accuracies reported in the case studies should only be seen as the *absolute minimum baseline*. Model comparison or prompt optimization is not the goal here; most prompts consist of simple 1–2 sentence instructions (see SI). Detailed expert prompts tend to yield better results. The accuracy rates are therefore not the focus, although they are reported—to illustrate tasks with (already present) potential for automation and to discuss how to incorporate them in subsequent statistical modeling.

Method details and statistical considerations

This section aims to situate and summarize some crucial aspects of applying a machine-assisted design. Practical aspects such as data preparation, data unitization, setting up an example LLM for machine annotation, and further statistical and transparent science considerations are all expanded upon in the SI, due to space constraints. See also Törnberg (2024) for further advice on setting up LLM annotators.

Conceptual method comparison. As an inherently mixed methods framework, MAQD incorporates advantageous aspects of qualitative and quantitative designs, while overcoming their respective limitations. A point-by-point method comparison can be found in the SI. The following summary is to situate the MAQD, but naturally, these archetypes may not capture every research scenario. Qualitative designs are deeply focused and can consider wider context, reception, societal implications, power relationships, and self-reflection, but the analyses are often difficult to generalize and estimate the uncertainty of, hard to replicate, or scale to large data. Quasi-quantifying designs try to retain this depth, and systematic coding supports replicability, but variable relationships and uncertainty remain impressionistic, while overconfidence in quantitative results without statistical modeling can lead to spurious results (see SI). Primarily quantitative methods are scalable (but only on countable data), relationships and their uncertainty can be estimated, and replication (or full reproduction, given data and procedures) is easier. Yet they may be seen as lacking the nuance and depth of the above. QD like feature analysis combines the benefits of qualitative analysis in the coding or transformation step with systematic quantification and uncertainty estimation. While coding involves subjectivity, it can be replicated, and statistical analyses reproduced. It can be applied in exploratory and confirmatory designs regardless of discipline. However, big data are a challenge due to the laborious human-annotation bottleneck. Machine assistance enables scaling up the QD while retaining all the benefits of qualitative analysis, systematic quantification, and replicability.

Even without machines, the QD provides a more systematic framework compared to approaches limited in these aspects. This includes applications to "small" interview or fieldwork data. Publishing the data, coding scheme, data processing, and statistical analysis code or instructions where possible fosters transparent, substantive scrutiny rather than conceptual or personal critique. QD is also a good alternative for getting at the "bigger picture" or themes, like in thematic analysis: unitizing and subsequent systematic quantitative modeling of theme distributions can only improve the induction of overarching themes, and unlike quasi-quantifying practices, also helps control for confounds, etc. It can also supplement discourse or content analysis research or approaches building on social semiotics (Halliday 1978). While the focus here is academic, the same approach could be used in analytics in business, marketing, media monitoring, communication management, etc. (Dell'Acqua et al. 2023; Mets et al. 2024).

Necessity of statistical modeling to avoid unintentional quasi-quantifying designs. The QD and MAQD only make sense if the inferred variables are subsequently modeled within a statistical framework to estimate the uncertainty of the estimates, whether they involve prevalence counts or complex multivariate relationships between the variables. In a hypothesis-driven research scenario, this typically also entails accounting for issues like confounding variables, interactions, and often enough, repeated measures (cf. Clark and Linzer 2015; Gries 2015; McElreath 2020; Winter 2020). None of these are exclusive to quantitative research, yet unfortunately often ignored in qualitative designs (some designs like qualitative comparative analysis do promise the evaluation of interactions, but only after data simplification). Estimating effect size (impossible in qualitative designs) helps to avoid sweeping yet spurious claims by showing how much variance is explained in social or cultural systems with typically many interacting variables. This complexity often renders simple pairwise tests inadequate, while more versatile models like multiple regression allow hypothesis testing while accounting for the aforementioned issues.

Inherent repeated measures are common in H&SS. Survey and interview data typically contain multiple responses, often with a variable number from multiple respondents. Literary or artistic examples may be sourced from multiple works from multiple authors from multiple further groupings, like eras or collections. Linguistic examples may be sourced from corpora (with several underlying sources) or elicited from multiple informants. Social media data often contain multiple data points per user. If the underlying grouping or hierarchical structure of a dataset is not accounted for, estimates can even reverse direction, known as Simpson's paradox (Kievit et al. 2013). This issue likewise applies in qualitative designs, but repeated measures cannot be systematically controlled there.

A necessary prerequisite for reliable analysis of initially unstructured or qualitative data (such as free-running text, film, etc.) is principled unitization, partitioning the data into meaningful divisions or units. In some cases the units are fairly obvious, e.g. paintings in a study about art, or sentences in syntax research; in others less so. It is important to make sure the units are comparable, large enough to be meaningful yet small enough to be succinctly analyzable (see the SI for extended discussion).

In summary, taking a systematic approach to both the initial data as well as the inferred variables is crucial to applying a QD or MAQD. There is not enough space here to delve into all possible pitfalls (and plenty of handbooks do), but a lack of control over these issues can easily lead to false, overestimated conclusions or even diametrically opposite results. Any quantitative claim, regardless of the label on the design, should be accompanied by an estimate of confidence or uncertainty. When making population-level claims from a sample, it is crucial to estimate reliability to avoid mistaking sampling noise for real differences, tendencies, or prevalences. In a crowdsourcing or machine-assisted scenario, quantifying annotation error is equally important. These issues are, however, not something that should be seen as complicating factors or requirements to make the researcher's life harder. On the contrary, it makes it easier: instead of worrying if an analysis is replicable and representative, the uncertainty can be estimated, enabling more principled interpretation and decision-making.

Incorporating classification error in statistical modeling. In a quantitizing research design, regardless if the annotation step is completed by humans or machines, inter-rater (dis)agreement should be accounted for in any subsequent operationalization of these new data, to avoid overconfident estimates and making Type I errors. It is far from atypical for (also human) annotation tasks to have only moderate agreement. This aspect is often ignored, even in QD applications like linguistic usage feature analysis, which usually otherwise strives for statistical rigor. As discussed in the "Introduction" section, no methodological element in this proposal is new on its own, including that of using LLMs to annotate data. What appears to be not yet widespread is the suggestion to systematically use expert knowledge to delegate coding, analysis, or annotation tasks to machines such as LLMs, while—importantly—also making sure the machine error rates are incorporated in statistical modeling and uncertainty estimates. Doing so enables using less than a perfect 100% accurate classifiers or annotators without fear of making overconfident or biased inferences down the line. Unless a closely comparable test set already exists, this will typically require a subset of data to be manually coded by human annotator(s) for evaluating the chosen machine(s).

Annotation error can be accounted for in several ways, including errors-in-variables (EIV) type regression models (with numerical variables), directly modeling measurement errors using

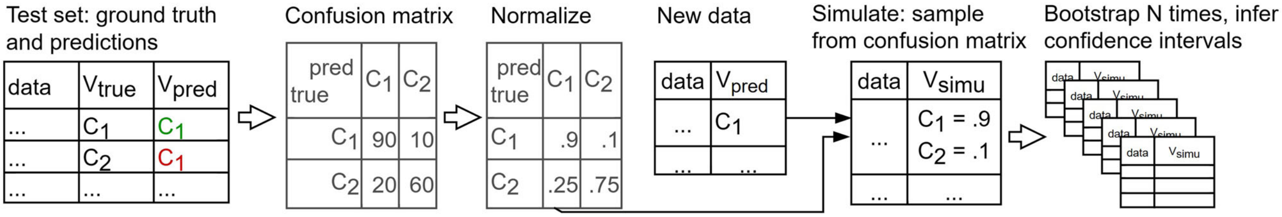

Fig. 2 A bootstrapping-driven pipeline for estimating the uncertainty in a machine-annotated categorical data variable *V*. The crucial component is the test set for comparing human expert annotation (ground truth) and machine predictions (or that of human coders). This provides an estimate of annotator accuracy and class confusion within the variable, which can then be used in bootstrapping the confidence intervals for the statistic of interest.

an MCMC or Bayesian approach (Carroll et al. 2006; Goldstein et al. 2008), or using prevalence estimation techniques (often referred to as “quantification” in ML; González et al. 2017). Distributional shift can be problematic but manageable (Guillory et al. 2021).

Keeping it simple, a bootstrapping approach is considered here which applies straightforwardly to exploratory and confirmatory scenarios and makes use of the rich error information from the confusion matrix between a ground truth test set and predictions or annotations. The procedure is simple, involving simulating the annotation procedure by sampling from the confusion matrix (see Fig. 2 for illustration; ground truth is rows, and predictions are columns).

1. Compute the confusion matrix m, of machine vs ground truth, for the variable of interest V which has a set of categorical levels or classes C.
2. For each class $i \in C$, normalize the count distribution of predictions against ground truth (“rows” of m) as probability distributions d_i.
3. Perform bootstrapping, creating N number of synthetic replicates of the data, where each predicted value V_j is replaced with a simulated value.
   3.1 For each case V_j with a value C_i, perform random weighted sampling from C using the corresponding d_i as weights.
   3.2 After simulating all values of V, perform the statistical analysis of interest (counts, prevalence, regression, etc.) on this synthetic dataset, and record the output(s).
4. Calculate ± confidence intervals for the statistic(s) of interest based on the estimate of the error (e.g. standard deviation) in the bootstrapped outputs.

For example, if the goal is to estimate the confidence of a percentage of class C_i in V, then the process is to perform bootstrapping on the raw new (classified or annotated) data some large number of times (e.g. 10,000), by sampling from the test set confusion matrix for each case in V, and calculating the statistic (percentage) in each replicate. Finally, calculate 95% confidence intervals via, e.g. normal approximation ($1.96 \cdot \sigma$). The intuition is: if the outputs of the (human or machine) annotator match with the ground truth 100%, then there will be no variance in the sampling either, and intervals will be ±0. The more confusion in the matrix, the more variance in the replicates, yielding wider confidence intervals. This is the simplest approach; potentially better and more robust procedures may be considered.

Results of case studies

For a class of machines to be considered for automation and augmentation in a machine-assisted design, their outputs should align closely enough with what human annotators or analysts would produce, while the remaining disagreement or error should be quantified and taken into account. Instructable LLMs have been shown to perform well in many domains and tasks. This section exemplifies the application of LLM-assisted MAQD through case studies in several languages and levels of difficulty as discussed in the “Introduction” section. Examples include explorative tasks, emulated analyses based on synthetic data, and a handful of replications of published research. Most are somewhat simplified examples, involving coding schemes of 1–2 variables, while real-world QD applications in complex cultural or social topics would typically involve several (in turn necessitating multivariable statistical modeling, see “Method details and statistical considerations” section). Two cloud service LLMs were used, gpt-3.5-turbo-0613 and gpt-4-0613 (OpenAI 2023), current at the time of conducting the studies, referred to as GPT-3.5 and GPT-4. The last, visual analytics example was updated to use the newer gpt-4o-2024-08-06 instead (current at the time of final revisions). Accuracy and Cohen’s kappa are used as evaluation metrics of machine performance in most case studies (see Table 1 in the SI for a bird’s eye view of the scores), as the focus here is agreement with human-annotated ground truth, rather than retrieval metrics. Accuracy illustrates empirical performance, while kappa adjusts for task difficulty by considering the observed and expected agreement given the number of classes.

Most of the case studies emulate or replicate only a segment of a research project pipeline, focusing on the applicability of LLMs in the quantitization step, but QD components of unitization, coding scheme, and (potential) quantification are described in each example. Further details on some case studies are found in the SI, along with the LLM prompts, due to journal space limitations. Many examples boil down to text-to-text, multi-class classification tasks—as much of H&SS is also concerned with various forms of classification, interpretation, and taxonomies, to enable predictions and discovering connections. Most tasks do not yield 100% agreement between human and machine annotations and analyses. This is not unexpected—less so because these machines have room to improve, and more because these are qualitative tasks requiring some degree of subjective judgment, where expecting 100% human agreement would not be realistic either. Human agreement rates could indeed be used to estimate an upper bound for machine accuracy in a task.

Example case: topic classification instead of latent topic modeling. In fields like digital humanities and computational social science, topic modeling is a commonly used tool to discover themes, topics, and their historical trends in media and literary texts, or to conduct “distant reading” (Jänicke et al. 2015; Moretti 2013). The bag-of-words LDA (Blei et al. 2003) is still widespread along with more recent sentence embedding-driven methods (Angelov 2020; Grootendorst 2022). These are all forms of soft or hard clustering: good for exploration, but suboptimal for confirmatory research. Historical or humanities researchers may often have hypotheses in mind, but they are difficult to test when they need to be aligned to ephemeral latent topics or clusters. Instead of such a tea leaves reading exercise, one could instead

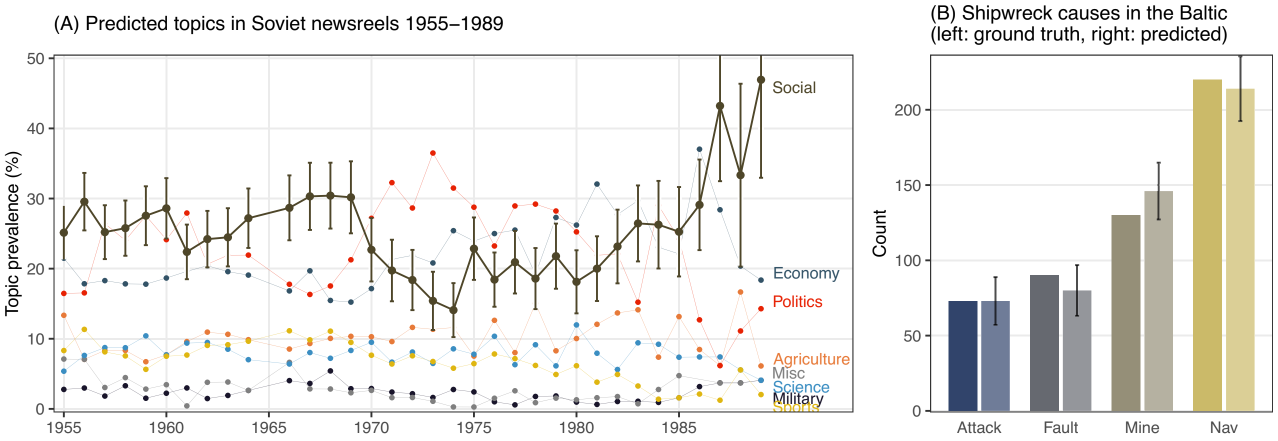

Fig. 3 Example results of applying a machine-assisted quantitizing approach to real-world data. A Zero-shot prediction of predefined topics in the corpus of Soviet newsreel synopses. The vertical axis shows yearly aggregate percentages. Bootstrapped confidence intervals are added to the trend of the Social topic. There is less data in the latter years, reflected in the wider intervals. **B** Wrecking causes of ships found in the Baltic Sea, mostly in Estonian waters, as annotated by experts based on field notes and historical documents (left), compared to zero-shot prediction of said categories based on the same data, with bootstrapped confidence intervals on the counts. Due to fairly good classification accuracy, the counts end up roughly similar.

classify texts by predefined topics of interest. Until recently, this would have, however, required training (or tuning an LLM as) a classifier based on labeled training data. Annotating the latter is time-consuming, so easy out-of-the-box methods like LDA have remained attractive (cf. Jacobs and Tschötschel 2019; Jelodar et al. 2019; Sherstinova et al. 2022).

With instructable LLMs, laborious training data compilation is no longer necessary, and topics can be predicted, instead of derived from clustering. Good prompt engineering is still necessary, but this is where qualitative scholars can be expected to shine. Ziems et al. (2023) worry that topic modeling may be challenging for transformer-based language models, given their context window size limitations, but even this is quickly becoming a non-issue. The GPT-4 version used here had a window size at 8000 tokens (about 6000 English words); the next iteration of GPT-4o models, released while this paper was in review, were already at 128,000, and Google's Gemini-1.5 models at 2 million (Kilpatrick et al. 2024). Also, longer texts can and indeed often should be split into smaller units (see SI).

The zero-shot topic classification approach is exemplified here using a dataset of Russian-language synopses of Soviet-era Moscow-produced newsreels from 1945 to 1992 (for details see Oiva et al. 2024). They summarize the story segments that comprise the roughly 10-minute weekly newsreel clips (12,707 stories across 1745 reels; each on average 16 words). As part of the collaboration cited above, an expert cultural historian defined, based on prior research and close viewing, a set of eight topics of interest, from politics to social and lifestyle to disasters (the latter was not encountered in the dataset; details and the longer topic definitions in the SI). An additional "miscellaneous" topic was defined in the prompt to subsume all other topics. Such an "everything else" or negative category is a naturally available feature in a zero-shot approach while requiring more complicated modeling in traditional supervised ML. We used an English-language prompt despite the data being in Russian, as this yielded better accuracy in early experiments (this makes sense given the dominant language of the model; cf. Wendler et al. 2024).

To summarize this in MAQD terms: the coding scheme consists of the categorical topic variable with nine levels, and the numeric variable of year (already present in the dataset). The unit is a story synopsis. The quantitization step consists of determining the primary topic of a given unit. The following quantification step involves aggregating topic counts as percentages, followed by regression analysis, with the machine-annotation uncertainty estimated using the aforementioned confusion matrix bootstrapping approach based on the expert test set. The final qualitative interpretation step could further contextualize the regression results with examples from the corpus, as in Oiva et al. (2024).

The expert annotated a test set of 100 stories for primary topic. Preliminary prompting experiments indicated that a single example per instruction prompt would yield the highest accuracy, which was 0.88 for GPT-3.5 (kappa 0.85; 0.84 and kappa 0.8 for GPT-4). This is of course the more expensive option when using cloud services like that of OpenAI that charge per input and output length. The cheaper option of batching multiple examples—preceded by an instruction prompt, requesting multiple output tags—seemed to reduce classification accuracy (this deserves further inquiry). While 88% accuracy is not perfect, recall that this is on a nine-class problem and a historical dataset rife with archaic terms and abbreviations that may not exist in the training data of a present-day LLM. The synopses, though short, sometimes contain different themes and topics, e.g. a featured tractor driver in an agricultural segment may also be lauded as a Soviet war hero or sports champion.

Following testing, GPT-3.5 was applied to the rest of the corpus of 12,707 stories, producing an estimate of news topic prevalence in the Soviet period (Fig. 3A). Among the trends, there is an apparent increase in the Social topic toward the end. However, the classifier is not 100% accurate; there are also fewer issues and therefore fewer data points toward the end. To test the trend, one can fit, for example, a logistic regression model to the period of interest (1974–1989), predicting topic by year (binomial, Social vs everything else). The model indicates a statistically significant effect of $\beta = 0.064$, $p < 0.0001$: each passing year multiplies the odds of encountering the Social topic by a factor of $e^{0.064} = 1.07$.

However, this is based on the predicted topics. A way to account for classifier error is bootstrapping, as discussed in the "Method details and statistical considerations" section: simulate the classification procedure by sampling from the confusion matrix, and rerun the statistical model across a large number of bootstraps. This yields distributions for each statistic of interest for inferring confidence intervals. Since the classifier here is fairly accurate, the 95% confidence interval around the log odds estimate is ±0.02, and for the p value, ±0.00002 (in other words, the upper bound still well below the conventional $\alpha = 0.05$). The same procedure was applied to infer intervals for the percent estimates in Fig. 3A.

Topic models are useful for exploration and discovery. Zero-shot topic and theme prediction is a viable alternative for testing confirmatory hypotheses about topical trends and correlations. If distributions of topics as in LDA are desired, lists or hierarchies of topic tags could be generated instead. In an additional but limited exploratory exercise, a sample of about 200 random synopses (about 8000 words in Russian, or 16k tokens) was fed into (gpt-3.5-turbo-16k), prompting it to come up with any number, and then also a fixed number of topics. These outputs were quite similar to the list initially produced by our expert historian by their admission.

Examples of historical event cause analysis and missing data augmentation. Detecting and extracting events, their participants, and causes from texts, including historical documents, is of interest in many fields in H&SS and information retrieval (Lai et al. 2021; Sprugnoli and Tonelli 2019). Ziems et al. (2023) experimented with applying LLMs to binary event classification and event argument extraction in English. Here, GPT-4 is applied to quantitizing the causes of shipwrecking events in an Estonian-language dataset of maritime shipwreck descriptions in the Baltic (a part of the Estonian state hydrography database, HIS 2023).

The unit of data is the database entry, containing a description of an incident based on various sources and fieldwork ($n = 513$, 1591–2006 but mostly twentieth century), including a primary cause of the wrecking variable. The already present initial coding scheme is a categorical variable with 54 unique values (a term or a phrase), inferred from the longer description by experts. This was simplified as a four-category variable: direct warfare-related (assaults, torpedo hits), mines, leaks or mechanical faults, and broadly navigational issues (such as getting caught on shallows or in a storm). The quantitization step consisted of inferring the category from the description texts, which range from short statements on sinking cause and location to longer stories such as this (here translated) example: *Perished en route from Visby to Norrkóping on the rocks of Västervik in April of 1936. After beaching in Gotland, Viljandi had been repaired, set sail from Visby to Norrkóping around mid-month. In a strong storm, the ship suffered damage to its rudder near Storkläppen, losing ability to steer. The injured vessel drifted onto the rocks of Sladó Island, and was abandoned. Local fisherman Ossian Johansson rescued two men from the ship in his boat. One of them was the ship's owner, Captain Sillen. Wreck sank to a depth of 12 meters.* (this is tagged as a navigation and weather-related wrecking).

GPT-4 accuracy was fairly high: the primary cause prediction matches with human annotation 88% (kappa 0.83; but, e.g. the "mine" class has a 100% recall). This is very good for a task with often multiple interacting causes and a sometimes arbitrary primary one, and on texts full of domain-specific terminology and Estonian maritime abbreviations. Figure 3B illustrates the quantification step, answering the question of the prevalence of wrecking causes, and how much a predicted distribution of causes would differ from an expert-annotated one, with bootstrapped confidence intervals on the counts (see "Method details and statistical considerations" section).

A quantitization task like this can also be used or viewed as missing data augmentation or imputation, where a variable with entirely or partially missing data is populated based on other variables. Real-world datasets in H&SS are often sparse, while applying multivariable statistics typically requires complete data. Another similar exercise was conducted using data from a Finnish-language broadcast management dataset, attempting to infer the missing production country entries of TV shows and films based solely on their (also very short) descriptions. The task was born out of real necessity when dealing with a highly sparse dataset. Despite its complexities and very open-ended nature (further details and examples in the SI), the results were promising, with GPT-4 yielding 72% accuracy against ground truth. These explorations demonstrate that already current LLMs are quite capable of completing inference even in contextual tasks that would otherwise have required manual work by domain experts.

Example of relevance classification with LLM-driven OCR correction in digitized newspaper corpora. A quantitizing design often requires first filtering and extracting relevant units from a larger pool of data. Digitization efforts of historical textual data such as newspapers and books have made large-scale, machine-assisted diachronic research feasible in many domains where this was not possible before. However, finding relevant examples from vast swathes of digitized data, more so if it is plagued by optical character recognition (OCR) errors, can be challenging. A part of the pipeline from a recent work (Kanger et al. 2022) is replicated here as an annotation exercise. The study sought to measure centurial trends in dominant ideas and practices of industrial societies focusing on the topic of nature, environment, and technology, based on digitized newspapers from multiple countries. Their pipeline for retrieving relevant examples for further analysis consisted of assembling a set of broad keywords, extracting passages where these occur, and estimating relevance ("nature" vs "human nature," etc.). The LDA topic modeling required cleaning and lemmatizing, and annotating the abstract topics for relevance. Such pipelines can be streamlined as a single operation on an LLM.

The authors kindly provided a human-annotated set of 99 excerpts for this exercise, extracted from a corpus using a keyword search. As such, the unit here is a context window around a keyword, the coding scheme is the categorical variable (nature-related or not), and quantitization consists of determining whether a given text snippet is nature-related. Being a filtering step, there is no further quantification here, but a real application could further model the co-occurrence of nature with other topics or variables or test hypotheses about temporal trends. While some parts of the sampled corpus were fairly readable, others contained OCR-distorted examples such as this:

principally to easing in » u ¿ allan consolidated bonds nine Issues Siorln « fallíand on'y two lssues ßalnl » 8 The lïttei Included the 3. per cent 1942 in which laigf pa'cek were bou.ht The Syd Ii, banks lollnqulshed a small pait of recent rlim Arünstnaturalleacilon in t. limited S, r of issues the main body of Indu- irai continued to find keen support.

This can however be "cleaned" with GPT-4: *principally to easing in Australian consolidated bonds; nine issues showing a fall and only two issues gaining. The latter included the 3 per cent 1942, in which large parcels were bought. The Sydney banks relinquished a small part of recent gains. As a natural reaction in the limited set of issues, the main body of industrial continued to find keen support.*

While such operations may suffer from LLM hallucination issues, we can test if this step degrades or improves the downstream classification results. The case turns out to be the latter. The prompt was to classify a given test input as mentioning "nature or environment in the biological natural world sense, including nature tourism, landscape, agriculture, environmental policy." Without the cleaning, GPT-3.5 gets 0.79 accuracy (0.49 kappa) and GPT-4: 0.9 (0.77). With cleaning, GPT-3.5 gets 0.82 (0.56) and GPT-4: 0.92 (0.82 kappa). This is another difficult task with limited and historical period-specific contexts, and more

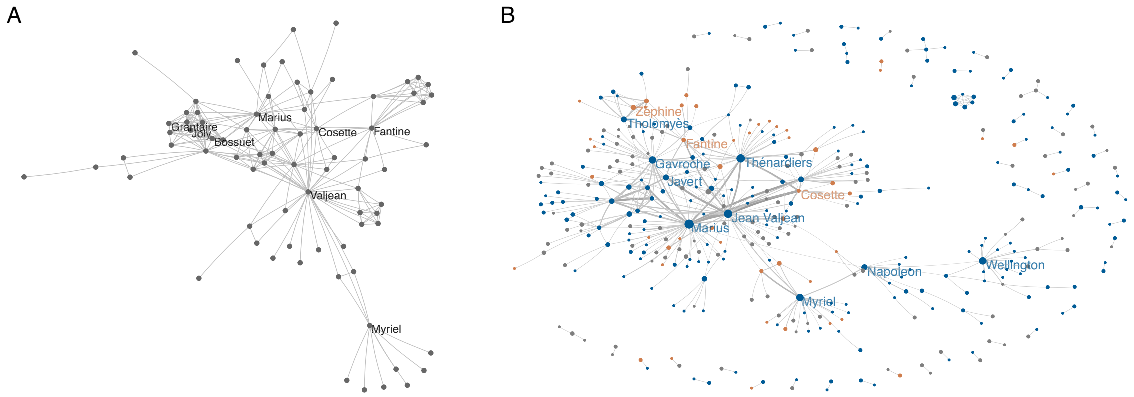

Fig. 4 Social networks of interacting characters in "Les Misérables" by Victor Hugo. A Manually constructed textbook example. **B** Automatically inferred using LLM (men are blue and women orange).

precise prompting might help. The results are promising however (in contrast to similar work, e.g. Boros et al. 2024). Using zero-shot or fine-tuned LLMs may well provide a simpler and faster alternative to complex processing and annotating pipelines such as this, as well as obviate the need for parameterizing and complex mathematical operations to make embedding vectors usable for that (cf. Sen et al. 2023), streamlining the application of MAQD approaches. The combination of an initial faster filter (keyword search, embedding similarity) with a follow-up generative LLM-based refining filter may well be a fruitful approach, as running an entire corpus through a cloud service LLM can be costly (and time-consuming, even if using a local model).

Linguistic usage feature analysis applications. As discussed in the "Introduction" section, ML and LLMs have found various uses in branches of linguistics. The application of LLMs in the usage feature analysis approach appears to be less explored. This case study replicates a part of the pipeline of a recent paper on linguistic value construction in eighteenth-century British English advertisement texts (Mulder et al. 2022). It focused on modifiers such as adjectives as a way of expressing appreciation in language and tested hypotheses about historical trends of both modifier usage and the advertised objects themselves. The paper details the process of developing the categories of "evaluative" (subjective) and "descriptive" modifiers via systematic annotation exercises and normalizing spelling in the historical texts (heterogeneous by nature and plagued by OCR errors) using edit distance, word embeddings, and manual evaluation. While cleverly utilizing computational tools, it is evident that no small amount of manual effort was expended. Most of such manual work can be streamlined and automated using zero-shot LLMs, including in low-quality OCR scenarios.

The replication here starts from the annotation, where the phrase units—such as *servants stabling* or *fine jewelry*—had already been extracted from a larger corpus (this is also a potential application for LLM-powered retrieval or filtering, as in the previous section). The coding scheme is the subjectivity variable (two levels); the quantitization step consists of assessing the relative subjectivity of the adjective in the phrase. A subsequent potential quantification step could aggregate those to compare against a time period variable as in the paper. GPT-4 agreement with expert human annotations is high (accuracy 0.94, kappa 0.89). For context, the paper reports the kappa agreement between the two researchers to have been at 0.84 in the first annotation iteration. This is not an easy task either and may require subjective decisions; e.g. *servants horse* is tagged descriptive (objective) yet *gentleman's saddle* as subjective in the dataset.

The same approach could also be used to automate or augment the quantitizing step in domains like grammar and syntax (cf. Beguš et al. 2023; Beuls and Van Eecke 2025; Qin et al. 2023; Szmrecsanyi et al. 2014). In another recent study (Karjus and Cuskley 2024) we experimented with automating semantic analysis, measuring linguistic divergence in US American English across political spectra. The unit: pairs of social media posts ($N = 320$). The coding scheme: an ordinal similarity variable from the DURel scheme (Schlechtweg et al. 2018); the quantitization: determining if a given target word or emoji (of 8 in total) is used in the same or different sense between them. GPT-4 achieved moderate correlation with the two human annotators ($\rho = 0.45$ and 0.6), struggling with the limited available context and context-specific emojis, underscoring the need to verify LLM accuracy against human judgment (see "Method details and statistical considerations" section and SI).

Example of social network inference from literary texts. Another potential application for MAQD is quantitizing data for network analysis. Here, an entire book was analyzed for character interactions: the unit is a chapter, the coding scheme consists of an open-ended annotation variable for any persons that interact (coded as pairs), and a gender variable for each character. The quantitization step involves finding instances of interacting pairs in a chapter (and marking character gender), and quantification consists of operationalizing the list of pairs as a weighted undirected network, where the nodes are characters and the links between them are weighted by the number of recorded interactions. Figure 4A depicts a character network manually constructed from "Les Misérables" by Victor Hugo, often used as a textbook example in network science and related fields. Figure 4B is a network of interacting characters inferred from the English version of the same book using GPT-3.5 according to the scheme above, and GPT-4 for estimating gender. The result may have some errors—some anomalous pairs like street names and unspecific characters ("people," etc.) required filtering, and better results may be achieved with a newer model. Still, the result is much richer than the smaller manual version, including non-plot characters discussed by Hugo in tangential sections. This demonstrates the applicability of LLMs to unitizing and quantitizing entire texts (the resulting units for the subsequent MAQD quantitative step being character pairs here), without requiring preprocessing with specialized models, for example, named entity recognition, syntactic parsing (cf. Elson et al. 2010). Newer models with larger context windows, yet unavailable when these studies were conducted, can also ingest entire books at once, but unitization may still be preferable for more controlled and fine-grained analysis.

Exemplary genre quantification in literature and film. Ziems et al. (2023) also discuss the computational analysis of literary themes, settings, emotions, roles, and narratives, and benchmark some of these tasks. The following is a replication of a recent study (Sobchuk and Šeļa 2024) which sought to compare computational methods of capturing and clustering thematic (genre) similarity of literary texts against manually assigned genres (Detective, Fantasy, Romance, Sci-Fi), using the Adjusted Rand Index (ARI; Hubert and Arabie 1985). Combinations of text embedding algorithms (bag-of-words, topic models, embeddings), parameters, preprocessing steps, and distance measures were evaluated. This replication illustrates both how instructable LLMs can be used as simpler alternatives to complex parameterized computational pipelines, and how to conceptualize longer texts such as books in a MAQD.

Given that ARI is equivalent to Cohen's kappa (Warrens 2008), the performance of an LLM, set up to classify genres, can be compared to the clustering results. The authors kindly shared their labeled 200-book test set. Rather than parsing entire books, 25 random passages (the unit here; 5000 in total) were sampled from each. The coding scheme consists of the genre variable, and GPT-3.5 was instructed to label each passage as one of the four genres. The best-performing parameter and model combination in Sobchuk and Šeļa (2024) used complex preprocessing, a doc2vec embedding (Le and Mikolov 2014), and cosine similarity for clustering. The preprocessing involved lemmatizing, NER and POS tagging for cleaning, and lexical simplification (using another word embedding). This yielded ARI 0.7. The simple zero-shot LLM approach here achieved a (comparable) kappa of 0.73 (0.8 accuracy) without any preprocessing, and only judging a subset of random passages per book (each book tagged simply by majority label; more data and a better model would likely improve results). Some genres were easier than others (Fantasy had 100% recall), while books combining multiple genres complicate things. These results reflect the result of the first case study: instead of clustering or topic modeling, zero-shot learning enables direct prediction and classification, and LLMs often obviate processing (see also Chaturvedi et al. 2018; Sherstinova et al. 2022).

Instead of labeling entire books with a single genre, on-demand classification like this can easily yield a more informative distribution of genres within each. Figure 5 illustrates an example application using two texts, P.K. Dick's "Do Androids Dream Of Electric Sheep?", and the adapted script of "Blade Runner" by H. Fancher and D. Peoples (the 1981 version). The script was unitized by scenes (but merging very short ones), and the book as equally sized chunks (note that the book is three times longer), and each unit was classified using GTP-3.5. Differences between the book and movie are revealed, the latter being more of a thriller with sci-fi and detective elements, while the book delves into various other topics. Both have detective elements in the first half and romantic ones in the middle. The one segment labeled "fantasy" does include the text "*At last a bird which had come there to die told him where he was. He had sunk down into the tomb world.*" This illustrates the potential of using zero-shot LLMs for scene analytics and quantifying fiction, either for exploratory or statistical modeling purposes, without needing to train specialized classifiers for each variable, while multimodal LLMs can add a visual dimension. Beyond this simple example, one could also consider running time or align book and adaption (cf. Yi et al. 2023), expand the coding scheme with mood, action, scenery, etc., and quantify their correlations.

A quantitizing approach to literary translation analysis. This section describes two experiments on the applicability of MAQD for analyzing literary translations. The unit of analysis is a sentence pair in both, and the quantitization consists of distilling differences between versions of a text into either an open-ended description or a quantifiable categorical variable. The data for the first task consists of the first paragraphs of G. Orwell's "1984" (until the "war is peace, freedom is slavery"), the English original and its Italian translation sentence-tokenized and aligned using the LLM-powered BERTalign (Liu and Zhu 2023), yielding 47 pairs. GPT-4 was prompted to examine and highlight any major lexical or stylistic differences in each pair. The outcome was evaluated by two native Italian-speaking literature scholars (see "Acknowledgments" section). Both concluded the alignment was correct and the inferences insightful, with no significant misinterpretations.

The second task involves a simulated test set of 100 English-Japanese sentence pairs (about a rabbit in a forest, children's fiction genre, GPT-4-generated). A total of 25 pairs match closely, but in 25 the rabbit is replaced with a bear in Japanese, in another 25 a moose is added, and in the last 25 the rabbit kisses his rabbit girlfriend, "redacted" in Japanese (emulating censorship scenarios not uncommon in the real world; cf. Inggs 2011). There are a few ways differences in text can be quantified. String edit distances are widely used in linguistics and NLP (Manning et al. 2008; Wichmann et al. 2010). Levenshtein distance is the sum of the optimal number of character additions, deletions, or substitutions to transform one string (e.g. word) to another. These work within the same language, but cannot capture synonymy or cross-linguistic differences. Machine translation algorithms or multilingual sentence embeddings can output a numeric similarity between two sentences in different languages, but a more fine-grained, interpretable metric might be useful.

The quantitization step consists of determining the value of such a "semantic edit distance" variable with four levels: no difference, addition, deletion, or substitution of contents or

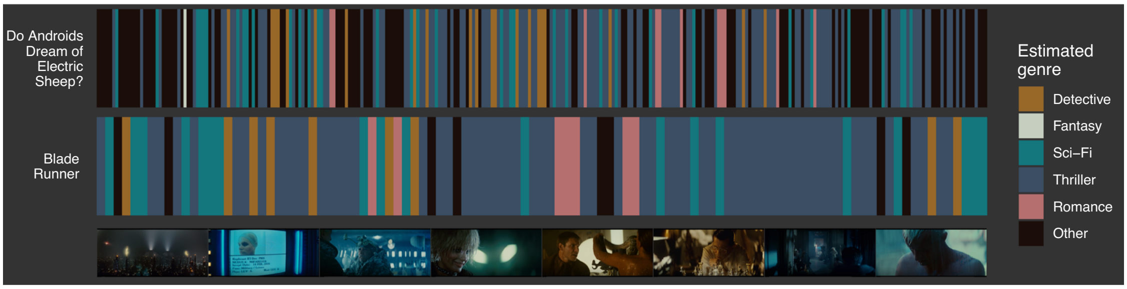

Fig. 5 Zero-shot classification of genre across one book and its film adaption, split into equally sized segments and scenes, respectively. Frames from the film are added for illustration. Differences and similarities become readily apparent and can provide a basis for follow-up qualitative or quantitative comparisons.

meaning. The dataset described above therefore has a ground truth total "distance" of 75/100. Further variables in real research could capture style, sentiment, author and translator metadata, etc., and associations between those could be statistically modeled. GPT-4 results on the current task are very good: 96% accuracy across the four classes (0.95 kappa; the simple sum of predicted difference classes is 74/100, just 1 off). Although small exercises, these examples show how combining multilingual LLM-powered auto-aligners with generative LLM-driven interpretation in a MAQD can enable scaling up the comparative analysis of translated, altered, or censored texts to larger text collections than a human researcher could manually process in a lifetime.

Further case studies: automating interview analysis and stance detection, and analyzing idea reuse, semantic change, and visual data. Given space limitations, the rest of the case studies are only briefly summarized here, with longer technical descriptions found in the SI.

Machine-quantitized interview analysis, opinion, and stance detection. Interview-based, phenomenological, and opinion studies across disciplines are often qualitative by label, yet can be observed at times giving in to the temptation of making quantitative claims about "often," "less," "many," etc. even when lacking systematic quantification or statistical modeling of the (un)certainty of such claims and the commonly repeated measures nature of the data (cf. Hrastinski and Aghaee 2012; Norman et al. 2021; Paasonen et al. 2023). A more systematic, e.g. a QD approach would be preferable (Banha et al. 2022). A simulated example is provided in the SI, illustrating a pipeline of qualitative coding of relevant passages (units) from interviews, followed by statistical modeling of the quantitized variables, to test an example hypothesis about distant learning, while controlling for age and repeated measures (multiple opinions per participant). The transformation of opinion into discrete variables can be outsourced to machines; in the example, GPT-4 detects the context-specific sentiment with 100% accuracy.

If the source of data is a larger text corpus or semi-structured interviews as above, it is often necessary to first detect and unitize relevant passages. In another upcoming paper, we will report on a cross-sector media monitoring collaboration with the Estonian Police and Border Guard Board, featuring an example of analyzing media texts for societal stances toward the police, combining keyword search with on-demand LLM-driven filtering and analysis. To briefly summarize here: we annotated 259 test sentences collected from a newspaper corpus, yielding 31 negative, 199 neutral, 19 positive, and 90 non-relevant sentences (e.g. containing police-related keywords but being about other countries or fiction or expressions like "fashion police"). In detecting relevance, GPT-3.5 had 76% accuracy; GPT-4 got 95% ($\text{kappa} = 0.9$, mean $\text{F1} = 0.9$). For stance, GPT-3.5 matched human annotations 78% while GPT-4 got 95% ($\text{kappa} = 0.88$; $\text{F1} = 0.92$), reflecting results of similar benchmarking exercises (Gilardi et al. 2023; Zhang et al. 2024), including in smaller languages (Mets et al. 2024; Rathje et al. 2024).

Text and idea reuse analysis. Political studies, history of ideas, cultural analytics, and science of science are among the disciplines interested in the spread and reuse of ideas and texts (see Chaturvedi et al. 2018; Gienapp et al. 2023; Linder et al. 2020; Salmi et al. 2021). Automated reuse detection may be based on keywords, latent embeddings, or hybrid approaches (Chaturvedi et al. 2018; Manjavacas et al. 2019). While detecting verbatim reprints of an entire news article or passage is often easy, tracking the reuse and spread of smaller units and abstract ideas is hard, more so if it crosses language boundaries. This exercise involved a simulated dataset of 100 social media post-like passages to test the feasibility of detecting the recurrence of a known pseudo-historical concept of "Russians are descendants of the Huns" (Oiva and Ristilä 2022). The quantization step: evaluating if a passage (the unit of data) contains the concept. In a real research scenario, this could be followed by a quantification step to model, e.g. the spread or change over time. GPT-4 is shown to complete the task with ease (virtually 100% accuracy), even after rephrasing, introducing OCR-like distortions, across translations between English and Russian, and combinations of all of the above (see the SI for details and examples).

Lexical semantic change detection. Unsupervised lexical change or shift detection is a task and research area in computational linguistics that attempts to infer changes in the meanings of words over time, typically based on large diachronic text corpora (Dubossarsky et al. 2019; Gulordava and Baroni 2011; Hamilton et al. 2016). A SemEval 2020 shared task (Schlechtweg et al. 2020) pitted numerous approaches against an annotated test of several languages and centuries of language change. The subtasks were: binary classification (which of the 27–38 test words per language have lost or gained senses between the given time periods?), and graded change detection (words ranked by change). Type-based embeddings were somewhat surprisingly still more successful in that task than contextual (BERT-like) models, which were later shown to require some task-specific engineering (Rosin and Radinsky 2022). The latter is the highest scoring approach (on task 2) on the same test set so far (Montanelli and Periti 2023). A semantic similarity, divergence, or change analysis may be practical to formulate in MAQD terms. Here the SemEval task is replicated with GPT-4 as the annotator. The unit: a pair of sentences containing a target word from a given language, sourced from the paired corpora provided in the task. Coding: the DURel scheme used for the original ground truth data of the task (cf. Schlechtweg et al. 2018). The quantitization: evaluate if the word occurs in the same or different meaning. Quantification: as per subtask, aggregation of change scores as binary decisions or a ranking. As further illustrated in the SI, this simple approach performs about as well as the best SemEval model in the binary task in English (70% accuracy) and outperforms it in the ranking task ($\rho = 0.81$ correlation vs 0.42, i.e. 2x improvement), as well as the more recent Rosin and Radinsky (2022) LLM-based result of 0.52. For German and Latin the results were not as good (see SI). This shows that a simple zero-shot machine-annotation approach can approach human accuracy in a difficult task like this, and can work on par or even surpass purpose-built complex architectures based on smaller LLMs or embeddings while requiring minimal effort to implement. This has also been further explored in recent works (Periti et al. 2024; Ren et al. 2024; Yadav et al. 2024) published between the preprinting and journal submission of the paper at hand.

Zero-shot sense inference for novel words. Another linguistic task that can benefit from QD thinking and machine automation is the semantic analysis of polysemous words, lexical innovations, or borrowings. The approach discussed above focuses on contrasting potentially diverged or changed meanings. Another is the feature analysis in corpus linguistics discussed in the "Introduction" section, where the componential meaning of (typically one or a group of) words may be analyzed via multiple fixed-level variables. The following is a more explorative task: inferring the meaning of previously unseen words. The test set of 360 sentences is again synthetic, split across three target senses (*bear*, *glue*, or *thief*; all represented by a placeholder nonce word), three languages (English, Turkish, Estonian), and two genres. The evaluation was iterative for testing purposes: the annotator was shown an increasingly larger set of shuffled examples of a given sense. The unit here is thus a set of 1–10 sentences, the coding scheme contains just the open-ended meaning variable, and GPT-4 was prompted to guess a one-word definition for the target placeholder given the context (without any further output constraints). A subsequent quantification step could in principle quantify, e.g. sense frequencies or trends over time or across genres and topics (cf. Karjus et al. 2020) to test linguistic hypotheses. The annotation results, further illustrated in the SI, are promising: while *bear* requires more examples (being confused with other large animals), GPT-4 can correctly infer the other two in all languages from just 3–4 examples. This shows both linguistics and applied lexicography can benefit from applying LLMs as machine assistants either in lieu or in conjunction with specialized models or human lexicographers (Lew 2023, 2024).

Visual analytics at scale using multimodal AI. The case studies above focused on the capabilities of LLMs in annotation and analysis of textual data. There is a clear direction toward multimodal models though, starting with vision. This includes the GPT models (OpenAI 2023), several other commercial frontier models, and open models like CogVLM, BLIP, or LLaVA, performing relatively well in complex visual tasks (Li et al. 2023; Wang et al. 2023b; Wu et al. 2023b). Various studies have explored the potential for using such models in media and cultural topics (Limberg et al. 2024; Lyu et al. 2023; Smits and Wevers 2023; Tang et al. 2024). The MAQD is data-agnostic and also suitable for analyzing visual data like paintings, photographs, book covers, film scenes, or museal artifacts. As with text, it is important to determine the units and coding scheme, follow systematic quantitization procedures, and apply suitable statistical modeling to the inferred variables. While additional visual case studies are not conducted, Fig. 6 depicts four examples of utilizing GPT-4o for image analytics. To avoid data contamination-driven results, all images were generated (using Stable Diffusion XL1.0), except for The Matrix lobby scene still, captured by the author. While these are all toy examples, scaling up such questions and inquiries to large datasets promises an unprecedented scale of analytics for fields like film studies, art history, visual anthropology, etc. Narrative descriptions of images may not be useful for quantification but illustrate the already available capacities of this class of models, which can reason about multilingual multimodal jokes (Fig. 6A) and produce coherent descriptions of realistic scenes as well as abstract visuals. Detecting and segmenting objects on images (Chen et al. 2023; Kirillov et al. 2023) or inferring art styles and aesthetics (Karjus et al. 2023; Mao et al. 2017) is nothing new as such, but vision-language models enable qualitative "reasoning" (Fig. 6A, B) and zero-shot classification and annotation (Fig. 6C).

Discussion

There is no substitute for expert human judgment, reasoning, and calibration, when it comes to designing, conducting and interpreting research. However, human time is a bottleneck. With suitable instructable machines, expertise can be scaled up to enable research on much larger volumes of data, enabling researchers to elevate to more representative sample sizes as well as ask new questions that were perhaps not yet feasible until recently. Humanities, social sciences, cultural analytics, and other neighbors of philosophy are very well positioned to leverage this opportunity, with long traditions in theory building, qualitative reasoning, and expert knowledge-driven analytics. The quantitizing design, exemplified by feature analysis in linguistics, provides a flexible and replicable framework. These competencies are crucial for its application, which requires solid questions or hypotheses, a well-aligned coding scheme, expert-annotated test sets for evaluation, and meaningful interpretation of the results of quantitative modeling of the quantitized variables.

Continuing to use potentially quasi-quantifying designs would seem difficult to justify, when more systematic options are available, employing transparent procedures, replicable annotation, and systematic quantification. Purely qualitative and conceptual research naturally has its place. However, applying qualitative designs in empirical scenarios where the actual (even if unstated) goal is quantification or extrapolation can lead to unintentional bad practices and spurious results. Where more data are available, limiting a study to tiny subsamples is no longer necessary, as previously time-consuming analysis tasks can be delegated to suitable machines. This is not to say LLMs or ML or AI should be applied to everything everywhere all at once. Automating annotation is rather an optimization problem between available resources, human time, and output quality. However, as shown above and in recent literature, using currently already available LLMs does not always decrease quality, and can

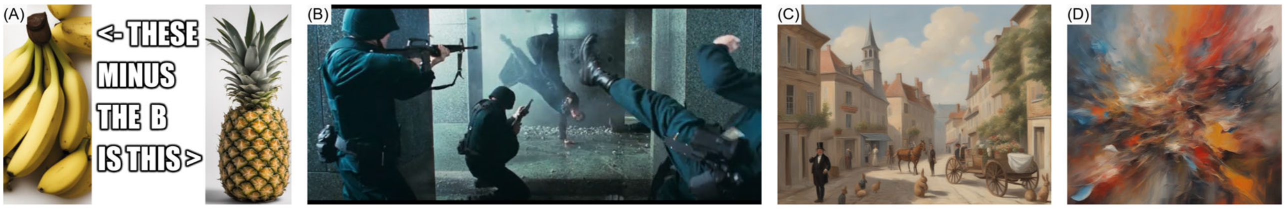

Fig. 6 Examples of visual analytics using the vision-capable GPT-4o. A Prompt: Explain this joke in one sentence. — Output: the joke plays on wordplay: removing the letter "B" from the word "bananas" leaves you with "ananas," which is another name for pineapple in many languages. **B** This is a still from a movie. Provide a list of likely genre, characters and their actions, location, mood, color palette, lighting, camera angle, shot type; all in variable: value format, short descriptions please. — Genre: Action, Sci-Fi. Characters: Protagonist (dodging), Armed agents (attacking). Location: Urban, interior, possibly a government or corporate building. Mood: Tense, high-stakes. Color Palette: Cool tones, mainly blues and greens. Lighting: Low key, dramatic shadows. Camera Angle: Low angle, emphasizing action. Shot Type: Medium-wide shot, capturing multiple characters and action. **C** Make a nested list of the most prominent objects and subjects in this painting, output as single-line compact minified JSON, use short one-word or descriptive camelCase labels only. — ["streetScene," ["buildings," "horse," "cart," "people," ["manWithTopHat," "womanInWhiteDress"], "bunnies," ["brownBunny," "blackBunny," "whiteBunny"], "trees," "clouds," "churchWithSpire"]]. **D** Describe the objective technical aspects of this painting like an art scholar, in one short concise paragraph. — This painting features a dynamic composition of swirling, intermingling colors with a strong emphasis on movement and texture. The color palette is dominated by warm tones of orange, red, and yellow, contrasted by cooler shades of blue and gray. The brushwork appears loose and expressive, with layers of paint blending seamlessly into one another, creating a sense of depth and fluidity. The overall form is abstract, with no discernible figures or objects, focusing on the interaction between the colors and the implied motion within the composition.

in some cases even improve upon human assistance (Gilardi et al. 2023; Mellon et al. 2024; Törnberg 2023).

Limitations and concerns. There are naturally inherent limitations to using such automation technology. Technological issues include model performance and inhibiting guardrails. Current models are typically limited to 1–2 modalities, but natural human communication is multimodal (cf. Rasenberg et al. 2022). This may well improve in the near future. One critique against using ML and LLMs to annotate data is that they can be unreliable or unreplicable and their outputs may be stochastic. Reasons include the nature of the underlying neural networks and loosely documented updates to cloud service LMMs. A related worry is that like all pretrained models, LLMs can inherit biases from their (often unknown) training data (Feng et al. 2023). However, as discussed elsewhere (Mellon et al. 2024; Törnberg 2023), these issues are not categorically unique to machines. They also apply to human analysts, crowd-worker annotators, research assistants, etc. Humans too are stochastic black boxes. Engaging in analytic tasks requiring subjective judgments and reasoning can propagate and amplify biases. The solution in qualitative (and mixed, quantitizing) research is to be mindful, acknowledge and attempt to control for this, and generally follow open science practices where possible (including preregistration) to enable transparent scrutiny and replicability.

Using open-source or open-weight LLMs based on well-documented training procedures and data would be preferable in that regard (cf. Liesenfeld et al. 2023). Running a fixed version of a local model can ease the replication issues that current cloud services may have if the model is public (in a persistent repository) or can be published along with the research. However, this is not always feasible (Palmer et al. 2024), such as at the time of conducting the case studies here, where the only LLMs capable of working with the smaller languages were the commercial ones. One might also criticize using LLMs for the fact that usage has costs, either cloud service fees or investments into capable local hardware. The ecological footprint of using LLMs has also been raised. Then again, arguably any research activity has costs and a footprint (see also Tomlinson et al. 2024), including hiring research assistants or crowdworkers—or using the most valuable resource, one's own time.

One way or another, LLMs are already being used in research, as discussed in the "Introduction" section. The cost and effort of running a typical "paper-sized" study has significantly decreased in many disciplines, especially those not requiring experimentation or primary data collection, while LLM-based tools like ChatGPT or Copilot also expedite writing. Anecdotally, the analytics for a typical feature-analytic linguistics paper involve annotating somewhere in the ballpark of 500 to 5000 examples, often sampled from a corpus; a PhD thesis about thrice that. Such a task can now be completed very fast using an LLM (at least at some level of quality). A discipline or a journal that allows itself to be flooded by low-effort, low-insight papers is bound to eventually erode its reputation while hindering real progress. Transparent practices and replicability have thus never been more important, and research evaluation might consider focusing less on volume and more on insight and intellectual contribution.

Future research and opportunities. While the case studies here aimed to cover a number of disciplines and task types, this contribution is by no means comprehensive in that regard. Using LLMs and multimodal models as zero-shot classifiers and inference machines holds potential for any field dealing with complex textual, visual, and otherwise qualitative data and products of human (and machine; Brinkmann et al. 2023) cultural evolution. Already currently available LLMs can be plugged into research pipelines for classification, analysis, data processing, and filtering. A single LLM prompt can often perform as well as complex multi-model preprocessing pipelines, which were of course necessary up until very recently—to the point of sometimes being research goals themselves (cf. Ash et al. 2024; Chaturvedi et al. 2018; Sherstinova et al. 2022; Sobchuk and Šeļa 2024). It may well make sense for researchers and groups to deploy custom model(s) on in-house hardware or private cloud, with fine-tuning for their domain and frequent use cases. It would be surprising if that would not become commonplace in the near future—but with great data processing power comes quantitative and statistical responsibility.

There are various other domains where generative models may be useful. One is experiments employing artificial languages or visual stimuli, as used in psychology and cognitive science (Galantucci et al. 2012; Karjus et al. 2021; Kirby et al. 2008; Tamariz and Kirby 2015), or cases where examples need to be anonymized (Asadchy et al. 2024), or open-ended responses analyzed. LLM-powered tools can be used to help build interfaces to run experiments. These are all tasks typically shared between a research team, but allocating some to machines means, for example, a cognitive scientist no longer needs to act as a full-stack developer, web designer, artist, and data annotator all in one.

Other areas may include law, educational sciences, or pedagogy. Empirical data like interviews, observations, practice reports but also laws and regulations can now be systematically analyzed at scale in a MAQD. In an educational setting, LLMs may be useful for assessment and other tasks (Baidoo-Anu and Owusu Ansah 2023; Kasneci et al. 2023). LLM-driven chatbots can also be used to conduct semi-structured interviews (Xiao et al. 2020). Another use case is potentially harmful, toxic, or triggering content, which can be delegated to a machine instead of horrifying a crowd-worker or RA.

One framework that relies on (machine-assisted) quantification of textual data is distant reading (Moretti 2013), typically culminating in interpreting word counts or topic distributions. Naturally, such representations are removed from the nuances of the content itself (the domain of "close reading"). One critique of Moretti-style distant reading (Ascari 2014) states that its reliance on predefined genre labels and "abstract models that rest on old cultural prejudices is not the best way to come to grips with complexity." In a MAQD, instead of relying on broad genre labels or abstract topic models, it is easy to model texts as distributions or sequences (of theory-driven units) at any level of granularity, including where the volumes of text would be unfeasible for close reading; machine reading instead of distant reading.

Time and efficiency gains. Ziems et al. (2023) suggest that the resources saved from allocating some tasks to LLMs would be put to good use by training expert human annotators or assistants. Indeed: let machines do repetitive labor and humans more interesting work. The time savings can be considerable. For example, the dataset of the first case study on newsreels features a modest 12,707 synopses totaling about 281k words. Assuming a reading speed of 184 wpm (words per minute; average for Russian-language text; Trauzettel-Klosinski et al. 2012), manual reading would be 19+ h of work, with annotation likely taking as much again; easily one working week—assuming the availability of a speaker of the language, able to interpret the historical context and the various abbreviations and references therein. Or a few hours on an LLM, yielding results very close to our expert with said qualifications.

The English translation of "Les Misérables" in the network example is about 558k words. Reading it would take 40 h (English average 228 wpm), taking notes of all interacting character pairs would likely double that; two weeks of work. Or a few LLM hours or even minutes. The Corpus of Historical American English (nineteenth–twentieth century; Davies 2010) is a common resource in linguistics. While NLP methods have been used to parse it for various ends, including the semantic change task discussed above, reading it all would take a human over 14 years (400M words; assuming 250 8 h-workdays per year). No scholar in their right mind would undertake this, so either small samples or aggregation is used. Meaningfully analyzing every single sentence therein using a pretrained LLM is very feasible.

The distant reading exercise purporting to launch a field of "culturomics" by analyzing "millions of books" (Michel et al. 2011) was based not on "books" but aggregated *n*-gram counts. The dataset was large though: even just the English segment (361B words) would be 13,000 years worth of reading. Processing a book dataset of that volume in a MAQD would likely take more than a few hours, but is doable with current LLMs, and would enable asking more meaningful questions than word frequencies can provide.

Conclusions

Recent benchmarking studies have demonstrated the applicability of pretrained LLMs to various data analytics and annotation tasks that until recently would have required human experts or complex computational pipelines. This contribution addressed the challenge of reliably making use of this automation and scaling potential in exploratory and confirmatory research, by advocating for a practical QD framework. Its applicability in tandem with current LLMs as instructable machine assistants was assessed in an array of multilingual case studies in humanities, social science, and cultural analytics topics. While using machine (and human) assistants does pose some risks, they can be largely mitigated through expert oversight, theory-driven analysis and unitization principles, rigorous quantitative modeling with error rate incorporation to avoid overconfident predictions, and generally transparent procedures and open science practices.

Data availability

The data and code used to run the analyses are available at https://github.com/andreskarjus/MachineAssisted.

Received: 2 May 2024; Accepted: 29 January 2025;
Published online: 28 February 2025

References

Acerbi A, Stubbersfield JM (2023) Large language models show human-like content biases in transmission chain experiments. Proc Natl Acad Sci 120(44):e2313790120

Aiyappa R, An J, Kwak H, Ahn Y-Y (2023) Can we trust the evaluation on ChatGPT? In: Proceedings of the 3rd workshop on trustworthy natural language processing (TrustNLP 2023). Association for Computational Linguistics, pp 47–54

Angelov D (2020) Top2Vec: distributed representations of topics. Preprint at https://doi.org/10.48550/arXiv.2008.09470

Asadchy Y, Karjus A, Mukhina K, Schich M (2024) Perceived gendered self-representation on Tinder using machine learning. Humanit Soc Sci Commun 11(1):1–11

Ascari M (2014) The dangers of distant reading: reassessing Moretti's approach to literary genres. Genre 47(1):1–19

Ash E, Gauthier G, Widmer P (2024) Relatio: text semantics capture political and economic narratives. Political Anal 32(1):115–132

Asher N, Bhar S, Chaturvedi A, Hunter J, Paul S (2023) Limits for learning with language models. In: Proceedings of the 12th joint conference on lexical and computational semantics (*SEM 2023). Association for Computational Linguistics, pp 236–248

Baidoo-Anu D, Owusu Ansah L (2023) Education in the era of generative artificial intelligence (AI): understanding the potential benefits of ChatGPT in promoting teaching and learning. J AI 7(1):52–62

Banha F, Flores A, Coelho LS (2022) Quantitizing qualitative data from semi-structured interviews: a methodological contribution in the context of public policy decision-making. Mathematics 10(19):3597

Beguš G, Dąbkowski M, Rhodes R (2023) Large linguistic models: analyzing theoretical linguistic abilities of LLMs. Preprint at https://doi.org/10.48550/arXiv.2305.00948

Bender EM, 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, FAccT '21. Association for Computing Machinery, pp 610–623

Beuls K, Van Eecke P (2025) Construction grammar and artificial intelligence. In: Fried M, Nikiforidou K (eds) The Cambridge handbook of construction grammar. Cambridge University Press, pp 543–571

Binz M, Schulz E (2023) Using cognitive psychology to understand GPT-3. Proc Natl Acad Sci 120(6):e2218523120

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022

Boros E, Ehrmann M, Romanello M, Najem-Meyer S, Frédéric K (eds) (2024) Post-correction of historical text transcripts with large language models: an exploratory study. In: Proceedings of the 8th joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature (LaTeCH-CLfL 2024). pp 133–159

Braun V, Clarke V (2012) Thematic analysis. In: APA handbook of research methods in psychology, vol 2: research designs: quantitative, qualitative, neuropsychological, and biological, APA handbooks in psychology. American Psychological Association, pp 57–71

Brinkmann L, Baumann F, Bonnefon J-F, Derex M, Müller TF, Nussberger A-M (2023) Machine culture. Nat Hum Behav 7(11):1855–1868

Buscemi A, Proverbio D (2024) ChatGPT vs Gemini vs LLaMA on multilingual sentiment analysis. Preprint at https://doi.org/10.48550/arXiv.2402.01715

Carroll RJ, Ruppert D, Stefanski, LA, Crainiceanu CM (2006) Measurement error in nonlinear models: a modern perspective, second edition. CRC Press

Chaturvedi S, Srivastava S, Roth D (2018) Where have I heard this story before? Identifying narrative similarity in movie remakes. In: Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: human language technologies, vol 2 (short papers). Association for Computational Linguistics, pp 673–678

Chen W, Li Y, Tian Z, Zhang F (2023) 2D and 3D object detection algorithms from images: a survey. Array 19:100305

Chi N, Malchev T, Kong R, Chi R, Huang L, Chi E et al. (2024) ModeLing: a novel dataset for testing linguistic reasoning in language models. In: Hahn M, Sorokin A, Kumar R, Shcherbakov A, Otmakhova Y, Yang J et al. (eds) Proceedings of the 6th workshop on research in computational linguistic typology and multilingual NLP. Association for Computational Linguistics, pp 113–119

Clark TS, Linzer DA (2015) Should I use fixed or random effects? Political Sci Res Methods 3(2):399–408

Creamer EG (2018) An introduction to fully integrated mixed methods research. SAGE Publications, Inc

Davies M (2010) The corpus of historical American English (COHA): 400 million words, 1810–2009

De Paoli S (2023) Performing an inductive thematic analysis of semi-structured interviews with a large language model: an exploration and provocation on the limits of the approach. Soc Sci Comput Rev 42(4):997–1019

de la Rosa J, Pérez Pozo A, Ros S, González-Blanco E (2023) ALBERTI, a multi-lingual domain specific language model for poetry analysis. Preprint at http://arxiv.org/abs/2307.01387

Dell'Acqua F, McFowland E, Mollick ER, Lifshitz-Assaf H, Kellogg K, Rajendran S. et al. (2023) Navigating the jagged technological frontier: field experimental evidence of the effects of AI on knowledge worker productivity and quality. Harvard Business School Technology & Operations Mgt. Unit Working Paper No. 24-013, The Wharton School Research Paper. Available at: https://ssrn.com/abstract=4573321

Demszky D, Yang D, Yeager DS, Bryan CJ, Clapper M, Chandhok S (2023) Using large language models in psychology. Nat Rev Psychol 2(11):688–701

Dentella V, Günther F, Leivada E (2023) Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias. Proc Natl Acad Sci 120(51):e2309583120

Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L (2023) QLoRA: efficient finetuning of quantized LLMs. Preprint at http://arxiv.org/abs/2305.14314

Dinh T, Zhao J, Tan S, Negrinho R, Lausen L, Zha S et al. (2023) Large language models of code fail at completing code with potential bugs. Preprint at http://arxiv.org/abs/2306.03438

Dubossarsky H, Hengchen S, Tahmasebi N, Schlechtweg D (2019) Time-out: temporal referencing for robust modeling of lexical semantic change. In:

Proceedings of the 57th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp 457–470

Elson D, Dames N, McKeown K (2010) Extracting social networks from literary fiction. In: Proceedings of the 48th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp 138–147

Erscoi L, Kleinherenbrink A, Guest O (2023) Pygmalion displacement: When humanising AI dehumanises women. Preprint at https://doi.org/10.31235/osf.io/jqxb6

Fan Y, Jiang F, Li P, Li H (2024) Uncovering the potential of ChatGPT for discourse analysis in dialogue: an empirical study. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). Association for Computational Linguistics, pp 16998–17010

Feng S, Park CY, Liu Y, Tsvetkov Y (2023) From pretraining data to language models to downstream tasks: tracking the trails of political biases leading to unfair NLP models. In: Proceedings of the 61st annual meeting of the Association for Computational Linguistics (volume 1: long papers). Association for Computational Linguistics, pp 11737–11762

Fetters MD, Curry LA, Creswell JW (2013) Achieving integration in mixed methods designs—principles and practices. Health Serv Res 48:2134–2156

Fofana F, Bazeley P, Regnault A (2020) Applying a mixed methods design to test saturation for qualitative data in health outcomes research. PLoS ONE 15(6):e0234898

Fonteyn L (2021) Varying abstractions: a conceptual vs. distributional view on prepositional polysemy. Glossa J Gen Linguist 6(1):90

Fudolig MI, Alshaabi T, Cramer K, Danforth CM, Dodds PS (2023) A decomposition of book structure through ousiometric fluctuations in cumulative word-time. Humanit Soc Sci Commun 10:187

Futrell R, Wilcox E, Morita T, Qian P, Ballesteros M, Levy R (2019) Neural language models as psycholinguistic subjects: representations of syntactic state. 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). Association for Computational Linguistics, pp 32–42

Galantucci B, Garrod S, Roberts G (2012) Experimental semiotics. Lang Linguist Compass 6(8):477–493

Gienapp L, Kircheis W, Sievers B, Stein B, Potthast M (2023) A large dataset of scientific text reuse in Open-Access publications. Sci Data 10(1):58

Gilardi F, Alizadeh M, Kubli M (2023) ChatGPT outperforms crowd workers for text-annotation tasks. Proc Natl Acad Sci 120(30):e2305016120

Glynn D (ed) (2010) Testing the hypothesis: objectivity and verification in usage-based cognitive semantics. In: Quantitative methods in cognitive semantics: corpus-driven approaches. De Gruyter Mouton, pp 239–269

Glynn D, Fischer K (2010) Quantitative methods in cognitive semantics: corpus-driven approaches. Walter de Gruyter, p 404

Goldstein H, Kounali D, Robinson A (2008) Modelling measurement errors and category misclassifications in multilevel models. Stat Model 8:243–261

González P, Díez J, Chawla N, Del Coz JJ (2017) Why is quantification an interesting learning problem? Prog Artif Intell 6(1):53–58

González-Bailón S, Lazer D, Barberá P, Zhang M, Allcott H, Brown T (2023) Asymmetric ideological segregation in exposure to political news on Facebook. Science 381(6656):392–398

Goodenough WH (1956) Componential analysis and the study of meaning. Language 32(1):195–216

Gries ST (2015) The most under-used statistical method in corpus linguistics: multi-level (and mixed-effects) models. Corpora 10(1):95–125

Gries ST, Divjak D (2009) A corpus-based approach to cognitive semantic analysis: behavioral profiles. In: Evans V, Pourcel S (eds) New directions in cognitive linguistics, human cognitive processing. John Benjamins Publishing Company, pp 57–75

Grootendorst, M (2022) BERTopic: neural topic modeling with a class-based TF-IDF procedure. Preprint at https://doi.org/10.48550/arXiv.2203.05794

Guillory D, Shankar V, Ebrahimi S, Darrell T, Schmidt L (2021) Predicting with confidence on unseen distributions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp 1134–1144

Gulordava K, Baroni M (2011) A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In: Proceedings of the GEMS 2011 workshop on geometrical models of natural language semantics. Association for Computational Linguistics, pp 67–71

Hagendorff T, Dasgupta I, Binz M, Chan SCY, Lampinen A, Wang JX et al. (2024) Machine psychology. Preprint at https://doi.org/10.48550/arXiv.2303.13988

Halliday MAK (1978) Language as social semiotic: the social interpretation of language and meaning. Edward Arnold, London, p 272

Hamilton WL, Leskovec J, Jurafsky D (2016) Diachronic word embeddings reveal statistical laws of semantic change. In: Proceedings of the 54th annual meeting of the Association for Computational Linguistics, ACL 2016, August 7–12, 2016, Berlin, Germany, vol 1: long papers, pp 1489–1501

Hau MF (2024) A practice-oriented framework for using large language model-powered chatbots. Acta Sociol https://doi.org/10.1177/00016993241264152

Hennessy S, Rojas-Drummond S, Higham R, Márquez AM, Maine F, Ríos RM (2016) Developing a coding scheme for analysing classroom dialogue across educational contexts. Learn Cult Soc Interact 9:16–44

Hesse-Biber SN (2010) Mixed methods research: merging theory with practice. Guilford Press, New York, p 242

HIS (2023) Hüdrograafia Infosüsteem https://his.vta.ee:8443/HIS/Avalik

Hrastinski S, Aghaee NM (2012) How are campus students using social media to support their studies? An explorative interview study. Educ Inf Technol 17(4):451–464

Hsieh H-F, Shannon S (2005) Three approaches to qualitative content analysis. Qual Health Res 15:1277–88

Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S. et al. (2021) LoRA: low-rank adaptation of large language models. Preprint at https://doi.org/10.48550/arXiv.2106.09685

Huang F, Kwak H, An J (2023) Is ChatGPT better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech. In: Companion Proceedings of the ACM web conference 2023, WWW '23 companion. Association for Computing Machinery, New York, NY, USA, pp 294–297

Hubert L, Arabie P (1985) Comparing partitions. J Classif 2(1):193–218

Huynh T, Hatton-Bowers H, Howell Smith M (2019) A critical methodological review of mixed methods designs used in mindfulness research. Mindfulness 10(5):786–798

Inggs J (2011) Censorship and translated children's literature in the Soviet Union: the example of the Wizards Oz and Goodwin. Target 23:77–91

Jacobs T, Tschötschel R (2019) Topic models meet discourse analysis: a quantitative tool for a qualitative approach. Int J Soc Res Methodol 22(5):469–485

Jänicke S, Franzini G, Cheema MF, Scheuermann G (2015) On close and distant reading in digital humanities: a survey and future challenges. In: Borgo R, Ganovelli F, Viola I (eds) Eurographics conference on visualization (EuroVis) – STARs. The Eurographics Association

Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y (2019) Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimed Tools Appl 78(11):15169–15211

Kane H, Kahwati L (2023) Mixed methods and qualitative comparative analysis. In: Tierney RJ, Rizvi F, Ercikan K (eds) International encyclopedia of education (fourth edition). Elsevier, Oxford, pp 581–587

Kanger L, Tinits P, Pahker A-K, Orru K, Tiwari AK, Sillak S (2022) Deep transitions: towards a comprehensive framework for mapping major continuities and ruptures in industrial modernity. Glob Environ Change 72:102447

Karjus A, Cuskley C (2024) Evolving linguistic divergence on polarizing social media. Humanit Soc Sci Commun 11(1):1–14

Karjus A, Blythe RA, Kirby S, Smith K (2020) Quantifying the dynamics of topical fluctuations in language. Lang Dyn Change 10(1):86–125

Karjus A, Blythe RA, Kirby S, Wang T, Smith K (2021) Conceptual similarity and communicative need shape colexification: an experimental study. Cogn Sci 45(9):e13035

Karjus A, Canet Solà M, Ohm T, Ahnert SE, Schich M (2023) Compression ensembles quantify aesthetic complexity and the evolution of visual art. EPJ Data Sci 12(1):1–23

Kasneci E, Sessler K, Küchemann S, Bannert M, Dementieva D, Fischer F (2023) ChatGPT for good? On opportunities and challenges of large language models for education. Learn Individ Differ 103:102274

Kievit RA, Frankenhuis WE, Waldorp LJ, Borsboom D (2013) Simpson's paradox in psychological science: a practical guide. Front Psychol 4:513

Kilpatrick L, Basu Mallick S, Kofman R (2024). Gemini 1.5 Pro 2M context window, code execution capabilities, and Gemma 2 are available today. Blog post, available at https://developers.googleblog.com/en/new-features-for-the-gemini-api-and-google-ai-studio/

Kirby S, Cornish H, Smith K (2008) Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language. Proc Natl Acad Sci 105(31):10681–10686

Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L et al. (2023) Segment anything. Preprint at http://arxiv.org/abs/2304.02643

Krippendorff K (2019) Content analysis: an introduction to its methodology, 4th edn. SAGE Publications, Inc

Lai VD, Nguyen MV, Kaufman H, Nguyen TH (2021) Event extraction from historical texts: a new dataset for black rebellions. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, pp 2390–2400

Le Q, Mikolov T (2014) Distributed representations of sentences and documents. In: Xing EP, Jebara T (eds) Proceedings of the 31st international conference on machine learning, vol 32. PMLR, pp 1188–1196

Lew R (2023) ChatGPT as a COBUILD lexicographer. Humanit Soc Sci Commun 10(1):704

Lew R (2024) Dictionaries and lexicography in the AI era. Humanit Soc Sci Commun 11(1):1–8

Li J, Li D, Savarese S, Hoi S (2023) BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. Preprint at http://arxiv.org/abs/2301.12597

Liesenfeld A, Lopez A, Dingemanse M (2023) Opening up ChatGPT: tracking openness, transparency, and accountability in instruction-tuned text generators. In: Proceedings of the 5th international conference on conversational user interfaces, CUI '23. Association for Computing Machinery, pp 1–6

Limberg C, Gonçalves A, Rigault B, Prendinger H (2024) Leveraging YOLO-World and GPT-4V LMMs for zero-shot person detection and action recognition in drone imagery. Preprint at https://doi.org/10.48550/arXiv.2404.01571

Linder F, Desmarais B, Burgess M, Giraudy E (2020) Text as policy: measuring policy similarity through bill text reuse. Policy Stud J 48(2):546–574

Liu L, Zhu M (2023) Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts. Digit Scholarsh Humanit 38(2):621–634

Lu C, Lu C, Lange RT, Foerster J, Clune J, Ha D (2024). The AI scientist: towards fully automated open-ended scientific discovery. Preprint at https://doi.org/10.48550/arXiv.2408.06292

Lund BD, Wang T, Mannuru NR, Nie B, Shimray S, Wang Z (2023) ChatGPT and a new academic reality: artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. J Assoc Inf Sci Technol 74(5):570–581

Lyu H, Huang J, Zhang D, Yu Y, Mou X, Pan J et al. (2023) GPT-4V(ision) as a social media analysis engine. Preprint at https://doi.org/10.48550/arXiv.2311.07547

Majumder BP, Li S, Ni J, McAuley J (2020) Interview: large-scale modeling of media dialog with discourse patterns and knowledge grounding. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). Association for Computational Linguistics, pp 8129–8141

Manjavacas E, Long B, Kestemont M (2019) On the feasibility of automated detection of allusive text reuse. In: Proceedings of the 3rd joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature. Association for Computational Linguistics, pp 104–114

Manning CD, Raghavan P, Schütze H (2008) Introduction to information retrieval. Cambridge University Press

Mao H, Cheung M, She J (2017) DeepArt: learning joint representations of visual arts. In: Proceedings of the 25th ACM international conference on multimedia, MM '17. Association for Computing Machinery, pp 1183–1191

McElreath R (2020) Statistical rethinking: a Bayesian course with examples in R and STAN, 2nd edn. CRC Press

Mellon J, Bailey J, Scott R, Breckwoldt J, Miori M, Schmedeman P (2024) Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale. Res Politics 11(1):20531680241231468

Mertens DM (2008) Transformative research and evaluation. Guilford Press, New York, p 417

Messeri L, Crockett MJ (2024) Artificial intelligence and illusions of understanding in scientific research. Nature 627(8002):49–58

Mets M, Karjus A, Ibrus I, Schich M (2024) Automated stance detection in complex topics and small languages: the challenging case of immigration in polarizing news media. PLoS ONE 19(4):e0302380

Miceli-Barone AV, Barez F, Konstas I, Cohen SB (2023) The larger they are, the harder they fail: language models do not recognize identifier swaps in python. In: Findings of the Association for Computational Linguistics: ACL 2023. Association for Computational Linguistics, pp 272–292

Michel J-B, Shen YK, Aiden AP, Veres A, Gray MK, Pickett JP (2011) Quantitative analysis of culture using millions of digitized books. Science 331(6014): 176–182

Mollick E (2024). Co-Intelligence. Penguin Random House

Montanelli S, Periti F (2023) A survey on contextualised semantic shift detection. Preprint at https://doi.org/10.48550/arXiv.2304.01666

Moretti F (2013) Distant reading. Verso Books

Morgan DL (1993) Qualitative content analysis: a guide to paths not taken. Qual Health Res 3(1):112–121

Motoki F, Neto VP, Rodrigues V (2023) More human than human: measuring ChatGPT political bias. Public Choice 198:3–23

Mulder AD, Fonteyn L, Kestemont M (2022) Linguistic value construction in 18th-century London auction advertisements: a quantitative approach. In: Proceedings of the computational humanities research conference 2022 (CHR 2022), vol 3. Antwerp, Belgium, 12–14 December, 2022, pp 92–113

Nadi F, Naghavipour H, Mehmood T, Azman, AB, Nagantheran, JA, Ting, KSK et al. (2024) Sentiment analysis using large language models: a case study of GPT-3.5. In: Bee Wah Y, Al-Jumeily OBE, Berry MW (eds) Data Science and Emerging Technologies. Springer Nature, pp 161–168

Nida EA (1979) A componential analysis of meaning: an introduction to semantic structures. In: Approaches to Semiotics [AS], vol 57. De Gruyter

Norman C, Wildman JM, Sowden S (2021) COVID-19 at the deep end: a qualitative interview study of primary care staff working in the most deprived areas of England during the COVID-19 pandemic. Int J Environ Res Public Health 18(16):8689

Novozhilova E, Mays K, Katz JE (2024) Looking towards an automated future: U.S. attitudes towards future artificial intelligence instantiations and their effect. Humanit Soc Sci Commun 11(1):1–11

O'Cathain A, Murphy E, Nicholl J (2010) Three techniques for integrating data in mixed methods studies. BMJ 341:c4587

O'Halloran KL, Tan S, Wignell P, Bateman JA, Pham D-S, Grossman M (2019) Interpreting text and image relations in violent extremist discourse: a mixed methods approach for big data analytics. Terrorism Political Violence 31(3):454–474

Oiva M, Mukhina K, Zemaityte V, Karjus A, Tamm M, Ohm T (2024) A framework for the analysis of historical newsreels. Humanit Soc Sci Commun 11(1):530

Oiva M, Ristilä A (2022) Mapping the pseudohistorical knowledge space in the Russian World Wide Web. In: Välimäki R (ed) Medievalism in Finland and Russia: twentieth- and twenty-first-century aspects. Bloomsbury, England, pp 57–71

Ollion É, Shen R, Macanovic A, Chatelain A (2024) The dangers of using proprietary LLMs for research. Nat Mach Intell 6(1):4–5

Ollion E, Shen R, Macanovic A, Chatelain, A (2023) ChatGPT for text annotation? Mind the Hype! Preprint at https://doi.org/10.31235/osf.io/x58kn

OpenAI (2023). GPT-4 technical report. Preprint at https://doi.org/10.48550/arXiv.2303.08774

Overos HD, Hlatky R, Pathak O, Goers H, Gouws-Dewar J, Smith K (2024) Coding with the machines: machine-assisted coding of rare event data. PNAS Nexus 3(5):165

Paasonen S, Sundén J, Tiidenberg K, Vihlman M (2023) About sex, open-mindedness, and cinnamon buns: exploring sexual social media. Social Media Soc 9(1):20563051221147324

Palaniyappan L (2023) Conversational agents and psychiatric semiology v3.0. Sistemi intelligenti, Rivista quadrimestrale di scienze cognitive e di intelligenza artificiale (2): 445–454. 10.1422/108140

Palmer A, Smith NA, Spirling A (2024) Using proprietary language models in academic research requires explicit justification. Nat Comput Sci 4(1):2–3

Parks L, Peters W (2023) Natural language processing in mixed-methods text analysis: a workflow approach. Int J Soc Res Methodol 26(4):377–389

Periti F, Dubossarsky H, Tahmasebi N (2024) (Chat)GPT v BERT Dawn of Justice for Semantic Change Detection. In: Graham Y, Purver M (eds) Findings of the Association for Computational Linguistics: EACL 2024. Association for Computational Linguistics, pp 420–436

Pilny A, McAninch K, Slone A, Moore K (2024) From manual to machine: assessing the efficacy of large language models in content analysis. Commun Res Rep 41(2):61–70

Qin C, Zhang A, Zhang Z, Chen J, Yasunaga M, Yang D (2023) Is ChatGPT a general-purpose natural language processing task solver? In: Bouamor H, Pino J, Bali K (eds) Proceedings of the 2023 conference on empirical methods in natural language processing. Association for Computational Linguistics, pp 1339–1384

Rasenberg M, Pouw W, Özyürek A, Dingemanse M (2022) The multimodal nature of communicative efficiency in social interaction. Sci Rep 12(1):19111

Rathje S, Mirea D-M, Sucholutsky I, Marjieh R, Robertson CE, Van Bavel JJ (2024) GPT is an effective tool for multilingual psychological text analysis. Proc Natl Acad Sci 121(34):e2308950121

Ren Z, Caputo A, Jones G (2024) A few-shot learning approach for lexical semantic change detection using GPT-4. In: Tahmasebi N, Montariol S, Kutuzov A, Alfter D, Periti F, Cassotti P et al. (eds) Proceedings of the 5th workshop on computational approaches to historical language change. Association for Computational Linguistics, pp 187–192

Rosin GD, Radinsky K (2022) Temporal attention for language models. In: Findings of the Association for Computational Linguistics: NAACL 2022. Association for Computational Linguistics, pp 1498–1508

Rytting CM, Sorensen T, Argyle L, Busby E, Fulda N, Gubler J et al. (2023) Towards coding social science datasets with language models. Preprint at https://doi.org/10.48550/arXiv.2306.02177

Salmi H, Paju P, Rantala H, Nivala A, Vesanto A, Ginter F (2021) The reuse of texts in Finnish newspapers and journals, 1771–1920: a digital humanities perspective. Historical Methods: J Quant Interdiscip Hist 54(1):14–28

Sandelowski M, Voils CI, Knafl G (2009) On quantitizing. J Mixed Methods Res 3(3):208–222

Schlechtweg D, McGillivray B., Hengchen, S., Dubossarsky, H., and Tahmasebi, N. (2020). SemEval-2020 Task 1: Unsupervised lexical semantic change detection. In: Proceedings of the fourteenth workshop on semantic evaluation. International Committee for Computational Linguistics, pp 1–23

Schlechtweg D, Schulte im Walde S, Eckmann S (2018) Diachronic Usage Relatedness (DURel): a framework for the annotation of lexical semantic change. In: Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: human language technologies, vol 2 (short papers). Association for Computational Linguistics, pp 169–174

Schleiger E, Mason C, Naughtin C, Reeson A, Paris C (2023) Collaborative intelligence: a scoping review of current applications. Preprint at Qeios, https://doi.org/10.32388/RZGEPB

Schreier M (2012) Qualitative content analysis in practice. Sage, London

Schrodt PA, Van Brackle D (2013) Automated coding of political event data. In: Subrahmanian V (ed) Handbook of computational approaches to counterterrorism. Springer, pp 23–49

Sclar M, Kumar S, West P, Suhr A, Choi Y, Tsvetkov Y (2023) Minding language models' (lack of) theory of mind: a plug-and-play multi-character belief tracker. In: Proceedings of the 61st annual meeting of the Association for Computational Linguistics, vol 1 (long papers). Association for Computational Linguistics, pp 13960–13980

Sen I, Quercia D, Capra L, Montecchi M, Šćepanović S (2023) Insider stories: analyzing internal sustainability efforts of major US companies from online reviews. Humanit Soc Sci Commun 10(1):309

Sherstinova T, Moskvina A, Kirina M, Zavyalova I, Karysheva A, Kolpashchikova E. et al. (2022) Topic modeling of literary texts using LDA: on the influence of linguistic preprocessing on model interpretability. In: 2022 31st conference of open innovations association (FRUCT), pp 305–312

Smits T, Wevers M (2023) A multimodal turn in Digital Humanities. Using contrastive machine learning models to explore, enrich, and analyze digital visual historical collections. Digit Scholarsh Humanit 38(3): 1267–1280

Sobchuk O, Šeļa A (2024) Computational thematics: comparing algorithms for clustering the genres of literary fiction. Humanit Soc Sci Commun 11(1):1–12

Sprugnoli R, Tonelli S (2019) Novel event detection and classification for historical texts. Comput Linguist 45(2):229–265

Szmrecsanyi B, Rosenbach A, Bresnan J, Wolk C (2014) Culturally conditioned language change? A multi-variate analysis of genitive constructions in ARCHER. In: Hundt M (ed) Late modern English syntax, studies in English language. Cambridge University Press, pp 133–152

Tamariz M, Kirby S (2015) Culture: copying, compression, and conventionality. Cogn Sci 39(1):171–183

Tang B, Lin B, Yan H, Li S (2024) Leveraging generative large language models with visual instruction and demonstration retrieval for multimodal sarcasm detection. In: Duh K, Gomez H, Bethard S (eds) Proceedings of the 2024 conference of the North American chapter of the Association for Computational Linguistics: human language technologies, vol 1 (long papers). Association for Computational Linguistics, Mexico City, Mexico, pp 1732–1742

Tanzer G, Suzgun M, Visser E, Jurafsky D, Melas-Kyriazi L (2024) A benchmark for learning to translate a new language from one grammar book. Preprint at https://doi.org/10.48550/arXiv.2309.16575

Tashakkori A, Teddlie C (2010) SAGE handbook of mixed methods in social & behavioral research. SAGE Publications, Inc

Taylor JET, Taylor GW (2021) Artificial cognition: how experimental psychology can help generate explainable artificial intelligence. Psychon Bull Rev 28(2):454–475

Tomlinson B, Black RW, Patterson DJ, Torrance AW (2024) The carbon emissions of writing and illustrating are lower for AI than for humans. Sci Rep 14(1):3732

Törnberg P (2023) ChatGPT-4 outperforms experts and crowd workers in annotating political Twitter messages with zero-shot learning. Preprint at https://doi.org/10.48550/arXiv.2304.06588

Törnberg P (2024) Best practices for text annotation with large language models. Sociologica, 18(2):67–85

Torrent TT, Hoffmann T, Almeida AL, Turner M (2023) Copilots for linguists: AI, constructions, and frames. In: Elements in construction grammar. Cambridge University Press, Cambridge

Trahan A, Stewart DM (2013) Toward a pragmatic framework for mixed-methods research in criminal justice and criminology. Appl Psychol Crim Justice 9(1):59–74

Trauzettel-Klosinski S, Dietz K, the IReST Study Group (2012) Standardized assessment of reading performance: the new international reading speed texts IReST. Investig Ophthalmol Vis Sci 53(9):5452–5461

Van Rooij I, Guest O, Adolfi F, de Haan R, Kolokolova A, Rich P (2024) Reclaiming AI as a theoretical tool for cognitive science. Comput Brain Behav 7:616–636

Vis B (2012) The comparative advantages of fsQCA and regression analysis for moderately large-N analyses. Sociol Methods Res 41(1):168–198

Wang C, Liu S, Li A, Liu J (2023a) Text dialogue analysis based ChatGPT for primary screening of mild cognitive impairment. Preprint at https://doi.org/10.1101/2023.06.27.23291884

Wang W, Lv Q, Yu W, Hong W, Qi J, Wang Y. et al. (2023b) CogVLM: visual expert for pretrained language models. Preprint at https://doi.org/10.48550/arXiv.2311.03079

Warrens MJ (2008) On the equivalence of Cohen's Kappa and the Hubert-Arabie adjusted Rand index. J Classif 25(2):177–183

Webb T, Holyoak KJ, Lu H (2023) Emergent analogical reasoning in large language models. Nat Hum Behav 7(9):1526–1541

Wendler C, Veselovsky V, Monea G, West R (2024) Do llamas work in English? On the latent language of multilingual transformers. In: Ku L-W, Martins A, Srikumar V (eds) Proceedings of the 62nd annual meeting of the Association for Computational Linguistics, vol 1 (long papers). Association for Computational Linguistics, pp 15366–15394

Wen-Yi AW, Adamson K, Greenfield N, Goldberg R, Babcock S, Mimno, D et al. (2024) Automate or assist? The role of computational models in identifying gendered discourse in US capital trial transcripts. Preprint at http://arxiv.org/abs/2407.12500

Wichmann S, Holman EW, Bakker D, Brown CH (2010) Evaluating linguistic distance measures. Phys A Stat Mech Appl 389(17):3632–3639

Winter (2020) Statistics for linguists: an introduction using R. Routledge, London

Wolk C, Bresnan J, Rosenbach A, Szmrecsanyi B (2013) Dative and genitive variability in late modern English: exploring cross-constructional variation and change. Diachronica 30(3):382–419

Wu T, Zhu H, Albayrak M, Axon A, Bertsch A, Deng W. et al. (2023a) LLMs as workers in human-computational algorithms? Replicating crowdsourcing pipelines with LLMs. Preprint at https://doi.org/10.1145/3706599.3706690

Wu W, Yao H, Zhang M, Song Y, Ouyang W, Wang J (2023b) GPT4Vis: what can GPT-4 do for zero-shot visual recognition? Preprint at https://doi.org/10.48550/arXiv.2311.15732

Xiao Z, Zhou MX, Liao QV, Mark G, Chi C, Chen W (2020) Tell me about yourself: using an AI-powered chatbot to conduct conversational surveys with open-ended questions. ACM Trans Computer-Hum Interact 27(3):15:1–15:37

Yadav S, Choppa T, Schlechtweg D (2024) Towards automating text annotation: a case study on semantic proximity annotation using GPT-4. Preprint at https://doi.org/10.48550/arXiv.2407.04130

Yi Q, Zhang G, Liu J, Zhang S (2023) Movie scene event extraction with graph attention network based on argument correlation information. Sensors 23(4):2285

Young SD, Jaganath D (2013) Online social networking for HIV education and prevention: a mixed-methods analysis. Sexually Transmitted Dis 40(2):162

Zhang W, Deng Y, Liu B, Pan S, Bing L (2024) Sentiment analysis in the era of large language models: a reality check. In: Duh K, Gomez H, Bethard S (eds) Findings of the Association for Computational Linguistics: NAACL 2024. Association for Computational Linguistics, pp 3881–3906

Zhu J-J, Jiang J, Yang M, Ren ZJ (2023) ChatGPT and environmental research. Environ Sci Technol 57(46):17667–17670

Ziems C, Shaikh O, Zhang Z, Held W, Chen J, Yang D (2023) Can large language models transform computational social science? Comput Linguist 1–53. 10.1162/coli_a_00502

Acknowledgements

The author would like to thank Mila Oiva for collaboration on the newsreels topic classification example which became an extended case study in this contribution, providing expertise in interpreting the additional results and feedback; Christine Cuskley for the collaboration on the Twitter paper, one component of which was also used here as an expanded example; Priit Lätti for providing a version of the maritime wrecks dataset, Daniele Monticelli and Novella Tedesco for providing expert evaluation in the English–Italian translation task, Laur Kanger and Peeter Tinits for discussions and for providing the test set for the historical media text filtering task, Oleg Sobchuk and Artjoms Šeļa for providing a test set for the literary genre detection task, and Tanya Escudero for discussions that led to expanding the literature review. Thanks for the useful discussions and manuscript feedback go to Vejune Zemaityte, Mikhail Tamm, Mark Mets, and Karmo Kroos.

Author contributions

The single author performed the research and wrote the paper. The author was supported by the CUDAN ERA Chair project for Cultural Data Analytics, funded through the European Union Horizon 2020 research and innovation program (Project No. 810961), and further supported by the Research Fund of the Tallinn University School of Humanities.

Competing interests

The author declares no competing interests.

Ethical approval

This article is based on publicly available textual data and simulated data, and does not contain any studies with human participants.

Informed consent

This article does not contain any studies with human participants.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-025-04503-w.

Correspondence and requests for materials should be addressed to Andres Karjus.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2025