An LLM-based pipeline for understanding decision choices in data analysis from published literature

ANONYMOUS AUTHOR(S)

Decision choices, such as those made when building regression models, and their rationale are essential for interpreting results and understanding uncertainty in an analysis. However, these decisions are rarely studied because tracing every alternatives considered by authors is often impractical, and reworking a completed analysis is generally of limited interest. Consequently, researchers must manually review large bodies of published analyses to identify common choices and understand how choices are made. In this work, we propose a workflow to automatically extract analytic decisions and their reasons from published literature using Large Language Models. Our method also introduces a paper similarity measure based on decision similarity and visualization methods using clustering algorithms. As an example, this workflow is applied to analyses studying the effect of particulate matter on mortality. This approach enables scalable and automated studies of decision choices in applied data analysis, providing an alternative to existing qualitative and interview-based studies.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; • **Information systems → Information retrieval**.

Additional Key Words and Phrases: large language models, analytic decision making in data analysis, document similarity

ACM Reference Format:

Anonymous Author(s). 2025. An LLM-based pipeline for understanding decision choices in data analysis from published literature. In *Proceedings of CHI Conference on Human Factors in Computing Systems (CHI'26).* ACM, New York, NY, USA, 23 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Data analysis is a complex and iterative process, and decisions are made at every stage of data analysis, from initial data collection, pre-processing, to modeling. One might expect well-trained researchers to make similar choices when faced with the same analytical task, yet evidence suggests otherwise. "Many-analyst" experiments show that independent analysts often arrive at markedly different conclusions, even when analyzing the same dataset to answer the same research question [8, 21, 68]. This variation in analytical decision-making, described by Gelman and Loken [20] as the "garden of forking paths," can undermine the quality and credibility of reported results and raise uncertainty in the findings.

A common approach to investigate uncertainty in data analysis decisions is sensitivity analysis, where researchers systematically vary key decisions in their analysis to assess the robustness of their findings. Multiverse analysis extends this idea by evaluating *all* plausible combinations of decision choices to examine how results vary across the full decision space [7, 63]. However, what an analyst considers reasonable may not reflect the full range of options used in practice. Even when a reasonable set of alternatives is tested, the stability shown by sensitivity analysis may be less relevant to other researchers with similar problems, who are often more interested in understanding the rationale

Manuscript submitted to ACM

behind decision choices. Ideally, decision-making in applied research can be studied by following experienced analysts throughout the entire analysis process to capture their reasoning. In reality, this is rarely feasible and not scalable. While individual studies may not capture the full range of decision choices used in practice, crowdsourcing decisions from a collection of studies on a shared theme creates a “many-analyst” setting that reveals how analysts make choices and justify them in practice. This process now has the possibility to be automated at scale, given recent advances in information extraction with Large Language Models (LLMs) [4, 19, 22, 24, 26, 28, 42, 47, 61, 65, 67].

In this work, we propose a new approach to studying data analysis decisions by automatically extracting decisions from scientific literature using LLMs. We develop a tabular schema to record decisions, automate the extraction process with LLMs, and introduce a new paper similarity measure based on decision similarity. This similarity measure can serve as a distance metric in dimension reduction methods to visualize papers according to their decisions. We apply this workflow to a set of 56 air pollution modeling studies that estimate the effect of particulate matter (PM2.5 or PM10) on mortality and hospital admissions, typically analyzed using Poisson generalized linear models (GLMs) or generalized additive models (GAMs). Analysis of the extracted decisions reveals common choices in this class of studies, including the use of smoothing methods on PM and weather variables and the temporal lags for time and weather variables. Multi-dimensional scaling on the paper similarity distance finds three distinct clusters corresponding to different smoothing methods: LOESS, natural spline, and smoothing spline. These findings align with the APHENA project [41], which synthesizes research from multiple studies in Europe and North America led by expert investigators. In this workflow, we also provide detailed documentation on the validation and standardization of LLM outputs. We outline the validation and standardization process, including the use of a developed Shiny application in R for reviewing decisions and the types of edits made through validation, the use of a secondary LLM to standardize reported choices of temporal lag decisions, and sensitivity analysis on reproducibility across runs and model providers.

In summary, the contribution of this work includes:

- A scalable and automated approach to study data analysis decisions through extracting of decisions from published scientific literature using LLMs,
- A new method to construct paper similarities based on decision choices and the semantic similarity of their rationales,
- Practices for validating and standardizing LLM outputs, including a shiny GUI tool for editing outputs, the use of secondary LLM for standardizing unstructured responses, and sensitivity analysis on reproducibility across runs and model providers,
- A data schema for recording decisions in data analysis in a tidy format, and
- A dataset of decisions, along with metadata, compiled from 56 studies in air pollution mortality modeling literature.

2 Related work

2.1 Analytic decision making in data analysis

Data analysis is a complex and iterative process [32–34] that involves multiple stages, including data collection, data cleaning, visualization, modeling, and communication. At each stage, analysts make decisions informed by domain practices, statistical knowledge, and the data. These decisions, such as which variables to include in a model, how to handle missing data, and how hyper-parameters are chosen, act as branching points in the analysis workflow. The full set of possible paths through these branching points forms what Gelman and Loken [20] describes as the “garden of forking paths". While one might expect well-trained researchers to make similar choices when facing similar decisions, empirical evidence suggests otherwise. "Many analyst experiments" show that independent research groups analyzing the same dataset to address the same research questions can arrive at widely different conclusions. For example, Silberzahn et al. [68] asks 29 groups of analysts to conduct an analysis to address the same research questions *whether soccer players with dark skin tone are more likely than those with light skin tone to receive red cards from referees*. Researchers reported an estimated effect size from 0.89 to 2.93 in odds ratio, with 21 unique combinations of covariates used among all 29 analyses. 70% of the teams found a statistically significant positive effect, while others didn't. This great discrepancy among researchers when performing data analysis tasks is also observed in other domains, for example, structural equation modeling [64], applied microeconomics [29], neuroimaging [8], and ecology and evolutionary biology [21].

Examples like the above illustrate how analytical decisions introduce uncertainty into data analysis. These uncertainties have been widely discussed in the literature, given their impact for policy recommendation [41] and domain applications e.g., fairness machine learning [70]. Through experiments, research has shown that analysts' decisions can lead to p-hacking and inflated effect size when not properly used [69, 74]. Hence, guidelines and checklists have been developed to recommend the best practices to guide statistical analysis. In medicine and biostatistics, pre-registration is a common practice to regulate analysts making decisions after seeing the data. Given the nuanced nature of data analysis, more work has examined how analysts make decisions in practice through interviews in both academia and industry. These studies include qualitative analysis of the decisions made [35, 49], interviews with data analysts about exploratory data analysis practice in industry [2, 38], and about how they consider alternatives in data analysis [48].

In addition to qualitative studies, software tools have been developed to help researchers account for alternatives and uncertainties and make informed decisions in data analysis. Examples include `Tea` [32], which supports general statistical analysis; `Tisane` [34], which guides choices in generalized linear mixed-effects models (GLMMs); and `MetaExplore` [36], which accounts for epistemic uncertainty (decision uncertainty) in meta-analysis. The `DeclareDesign` package [7] proposes the MIDA framework for researchers to declare, diagnose, and redesign their analyses to account for uncertainties of reporting the statistic of interest. Multiverse analysis proposes a different method to allow researchers to evaluate *all* plausible combinations of decision choices to examine how results vary in the full decision space. Work has been done on the software tools to support multiverse analysis [25, 63] and visualization of multiverse results [50], and debugging tools [23].

2.2 Automatic information extraction with LLMs

In natural language processing, information extraction is a task focused on extracting structured information from unstructured text. Earlier approaches in information extraction tasks relied on rule-based systems and regular expressions. More recent advances, including conditional random fields [45], word embeddings such as word2vec [54], and transformer-based architectures like BERT [16], have led to the current use of LLM to extract information with prompts. Using LLMs to extract unstructured text offers the advantage of automating the process at scale. Applications have been seen in epidemiology data [26], scientific literature [42], clinical data [19, 22, 28, 67], chemistry knowledge [65], and polymer science [24], climate extreme impact [47], phenotypes [4], and material properties [61]. An easier task in information extraction is called Named Entity Recognition (NER) to identify short span information (1-4 tokens) like person names and locations from unstructured text [57]. An example of this is extracting patients' information and vitals in clinical data. Extracting decisions from published literature is a more general task than NER, since justification of a decision typically spans more than just a few words. Our task also requires linking information across sentences, sometimes sections, to correctly identify the variables a decision refers to.

2.3 Visualization on scientific literature

With the growing volume of scientific publications and the difficulty of navigating the literature, there is an increasing interest in developing systems to visualize and recommend scientific papers. These systems link papers based on their similarity and relevance, typically determined by keywords [30], citation information [14], e.g., citation list and co-citation, or combinations with other relevant paper metadata [6, 15, 18, 27], e.g., author and title. Recent approaches incorporate text-based information using topic modeling [1], argumentation-based information retrieval [71], and text embedding [58]. While metadata and high-level text-based information are useful for finding relevant papers, researchers also need tools that help them *make sense* of the literature rather than simply *locating* it. In applied data analysis, one interest is to understand how studies differ or align in their decision choices. Capturing the decision choices and reasons that justify the choices from analyses enables the calculation of similarity among papers and can be piped into dimension reduction methods and visualization for a global view of analysis practice in the field or recommend similar papers based on decision similarities.

3 Methods

In this section, we present the workflow for extracting decisions from published literature using LLMs. We first describe the data structure for recording decisions, followed by the four main steps in the workflow: 1) automatic extraction of decisions from literature with LLMs, 2) validation and standardization of LLM outputs, 3) calculation of paper similarity, and 4) visualization of paper similarity using clustering or dimension reduction methods. The section concludes with an illustration summarizing the workflow.

3.1 Record decisions in data analysis

In the study of the health effects of outdoor air pollution, one area of interest is the association between short-term, day-to-day changes in particulate matter air pollution and daily mortality counts. This question has been studied extensively by researchers across the globe, and it serves to provide scientific evidence in the US to guide public policy on setting the National Ambient Air Quality Standards (NAAQS) for air pollutants. While individual modeling choices vary, these studies often share a common structure: they adjust for meteorological covariates, such as temperature and humidity, include lagged variables to account for temporal correlations, and estimate the effect size by city or region before pooling the results with random effect. This naturally forms a "many-analyst" experiment setting to analyze decisions in air pollution mortality modelling.

Consider the following excerpt from Ostro et al. [59] modeling the association between daily counts of mortality and ambient particulate matter (PM10):

> Based on previous findings reported in the literature (e.g., Samet et al. 2000), the basic model included a smoothing spline for time with 7 degrees of freedom (df) per year of data. This number of degrees of freedom controls well for seasonal patterns in mortality and reduces and often eliminates autocorrelation.

This sentence encodes the following components of a decision:

- **variable**: time
- **method**: smoothing spline
- **parameter**: degree of freedom (df)
- **reason**: Based on previous findings reported in the literature (e.g., Samet et al. 2000); This number of degrees of freedom controls well for seasonal patterns in mortality and reduces and often eliminates autocorrelation.

- **decision**: 7 degrees of freedom (df) per year of data

This decision can be recorded in a tabular format following the tidy data principle [75], which states that each variable forms a column and each observation forms a row. For our purpose, each row represents a decision made in a paper, and an analysis often includes multiple decisions. We extract the original text in the paper, without paraphrasing or summary. The decision above is a parameter choice of a statistical method applied to the variable *time*. A data analysis may also include other types of decisions, such as temporal or spatial ones, for example, the choice of lagged exposure for certain variables or whether the model is estimated collectively or separated for individual locations. These decisions don't have a specific method or parameter fields, but should still include variable, type (spatial or temporal), reason, and decision fields.

Given the writing style of authors, multiple decisions may be combined in one sentence, and certain fields may be omitted. Consider a different excerpt from Ostro et al. [59]:

> Other covariates, such as day of the week and smoothing splines of 1-day lags of average temperature and humidity (each with 3 df), were also included in the model because they may be associated with daily mortality and are likely to vary over time in concert with air pollution levels.

This sentence contains four decisions: two for temperature (the temporal lag and the smoothing spline parameter) and two for humidity, and should be structured as separate entries:

Paper	ID	variable	method	parameter	type	reason	decision
ostro	1	temperature	smoothing spline	degree of freedom	parameter	3 degree of freedom	NA
ostro	2	relative humidity	smoothing spline	degree of freedom	parameter	3 degree of freedom	NA
ostro	3	temperature	NA	NA	temporal	1-day lags	NA
ostro	4	relative humidity	NA	NA	temporal	1-day lags	NA

Notice in the example above, the reason field is recorded as NA. This is because the stated reason ("and are likely to vary over time in concert with air pollution levels") only supports the general inclusion of temporal lags but does not justify the specific choice of 1-day lag over other alternatives, e.g., 2-day average of lags 0 and 1 or single-day lag of 2 days. Similar scenarios can happen when a direct decision choice is missing, but a reason is provided, as in Katsouyanni et al. [40]:

> The inclusion of lagged weather variables and the choice of smoothing parameters for all of the weather variables were done by minimizing Akaike's information criterion.

3.2 Extract decisions automatically from literature with LLMs

Manually extracting decisions from published papers is labor-intensive and time-consuming. With LLMs, it is now possible to automatically extract this type of information by supplying a set of PDF documents and a prompt for instruction. Text recognition from PDF documents relies on Optical Character Recognition (OCR) to convert scanned images into machine-readable text, a capability currently offered by Antropic Claude and Google Gemini. In the prompt, we assign the LLM a role as an applied statistician and instruct it to extract decisions from the PDF in the format described in Section 3.1 and write the output in a JSON block in a markdown file. We also provide a set of instructions and examples on the possibility of missing reason and decision fields as discussed in Section 3.1. Prompt engineering techniques [13, 78] are used to optimize the prompt, and the full prompt used in this work is provided in the Appendix. We use the `chat_PROVIDER()` functions from the `ellmer` package [77] in R to obtain the output.

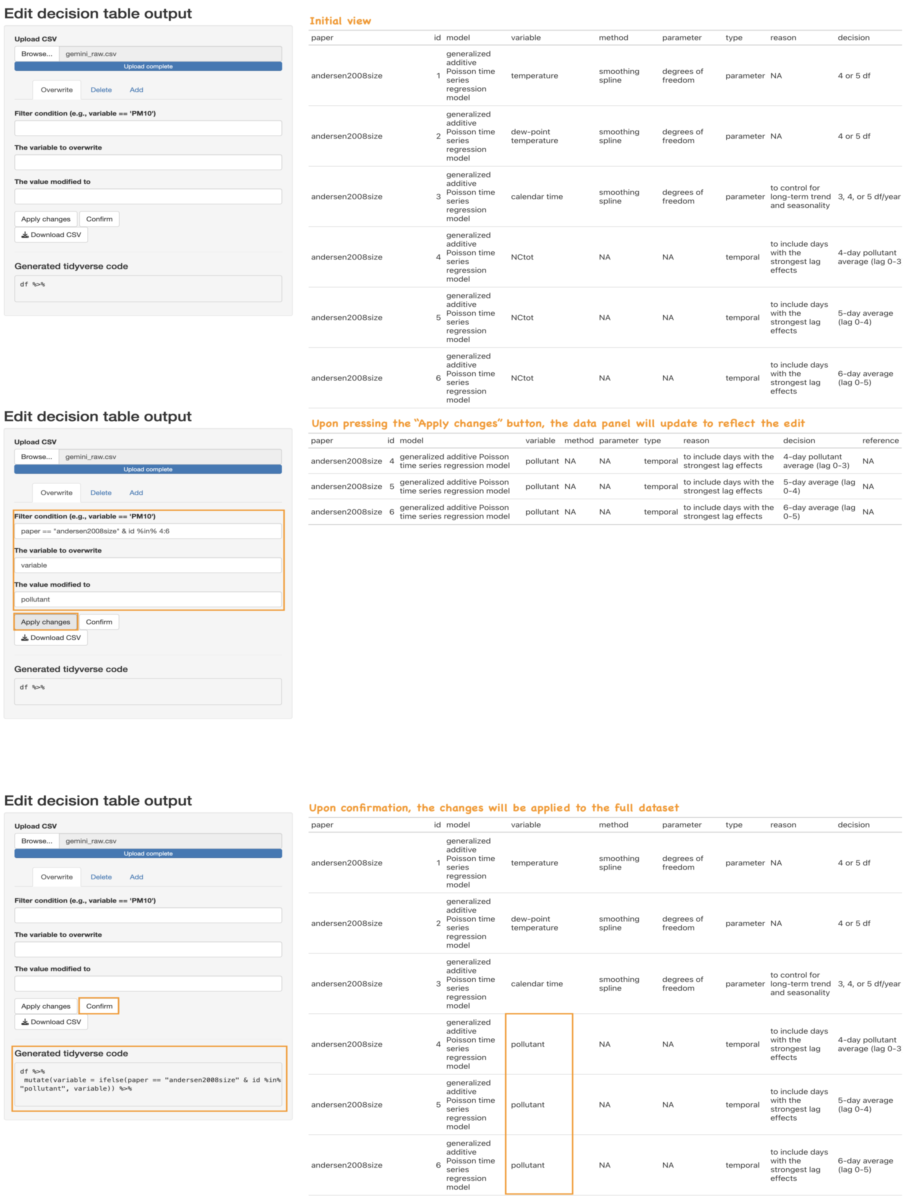
3.3 Validate and standardize LLM outputs

The LLM outputs need to be validated and standardized before further analysis. Validation focuses on ensuring the extracted decisions are correct, while standardization ensures different expressions of the same variable are standardized into the same expression. For example, the expressions *mean temperature*, *average temperature*, and *temperature* all refer to the same variable and are standardized to *temperature*. To help with the validation and standardization process, we developed a Shiny application, which provides an interactive interface for users to review and edit the LLM outputs. The Shiny application takes an input of a CSV file that contains the extracted decisions and allows users to perform three types of edits: 1) *overwrite* – modify the content of a particular cell, 2) *delete* – remove an irrelevant decision, and 3) *add* – manually enter a missing decision. Figure 1 illustrates the *overwrite* action for standardizing the variable *NCtot* (number concentration of particles <100 nm in diameter) to *pollution*. The user enters a predicate function in the filter condition box on the left panel, and the filtered data will appear interactively on the right panel. The user can then specify the variable to overwrite and the new value. The corresponding cells on the right panel will be updated. This change needs to be confirmed by pressing the "Apply changes" button to update to the full dataset. The corresponding `tidyverse` [76] code will then be generated on the left panel to be included in an R script, and the edited table can be downloaded for future analysis.

3.4 Calculate paper similarity and visualization

Once the output has been extracted and validated, these decisions can be treated as data for further analysis. Apart from exploratory data analysis, we propose a paper similarity measure to compare how similar decisions are between paper pairs. A decision is considered comparable between a paper pair if the two papers share the same variable and decision type, e.g., a parameter decision on temperature. Three factors are considered in calculating the similarity between two matched decisions: 1) whether the two decisions are similar, 2) whether the reasons for the decisions are similar, and 3) for parameter type decisions, whether the statistical methods used are the same. Method and choice similarity indicate the same decision being made in the analysis, whereas a similar reason reflects a shared principle for making the choice, even when the choices themselves may differ due to differences in the underlying data. For reasons and choices, we first obtain the text embedding for all the choices and reasons, and calculate the cosine similarity between the matched reason and decisions from the language model `BERT` using the `text` package [43] in `R`. For methods, we encode them as a binary variable: 1 if the two papers used the same method, and 0 otherwise, because semantic similarity cannot fully capture the difference between statistical methods, e.g., the difference between smoothing spline and natural spline is not well represented by the textual difference of "smoothing" and "natural". The paper similarity is then computed as the average decision similarity scores across all the matched methods, decisions, and reasons.

Although paper similarity can be calculated based on all available matched decisions, care should be taken for pairs with only a small number of matches. This can happen because two papers focus on different variables or some decisions have missing choices or reasons (discussed in Section 3.1). In practice, users may decide to focus on a set of decisions shared among papers or on papers that report a minimal number of shared decisions when calculating paper similarity.



Edit decision table output

Upload CSV: gemini_raw.csv

Filter condition (e.g., variable == 'PM10'):

The variable to overwrite:

The value modified to:

Generated tidyverse code

```
df %>%
```

Initial view

paper	id	model	variable	method	parameter	type	reason	decision
andersen2008size	1	generalized additive Poisson time series regression model	temperature	smoothing spline	degrees of freedom	parameter	NA	4 or 5 df
andersen2008size	2	generalized additive Poisson time series regression model	dew-point temperature	smoothing spline	degrees of freedom	parameter	NA	4 or 5 df
andersen2008size	3	generalized additive Poisson time series regression model	calendar time	smoothing spline	degrees of freedom	parameter	to control for long-term trend and seasonality	3, 4, or 5 df/year
andersen2008size	4	generalized additive Poisson time series regression model	NCtot	NA	NA	temporal	to include days with the strongest lag effects	4-day pollutant average (lag 0-3
andersen2008size	5	generalized additive Poisson time series regression model	NCtot	NA	NA	temporal	to include days with the strongest lag effects	5-day average (lag 0-4)
andersen2008size	6	generalized additive Poisson time series regression model	NCtot	NA	NA	temporal	to include days with the strongest lag effects	6-day average (lag 0-5)

Edit decision table output

Upload CSV: gemini_raw.csv

Filter condition (e.g., variable == 'PM10'): paper == "andersen2008size" & id %in% 4:6

The variable to overwrite: variable

The value modified to: pollutant

Generated tidyverse code

```
df %>%
```

Upon pressing the "Apply changes" button, the data panel will update to reflect the edit

paper	id	model	variable	method	parameter	type	reason	decision	reference
andersen2008size	4	generalized additive Poisson time series regression model	pollutant	NA	NA	temporal	to include days with the strongest lag effects	4-day pollutant average (lag 0-3)	NA
andersen2008size	5	generalized additive Poisson time series regression model	pollutant	NA	NA	temporal	to include days with the strongest lag effects	5-day average (lag 0-4)	NA
andersen2008size	6	generalized additive Poisson time series regression model	pollutant	NA	NA	temporal	to include days with the strongest lag effects	6-day average (lag 0-5)	NA

Edit decision table output

Upload CSV: gemini_raw.csv

Filter condition (e.g., variable == 'PM10'):

The variable to overwrite:

The value modified to:

Generated tidyverse code

```
df %>%
 mutate(variable = ifelse(paper == "andersen2008size" & id %in%
"pollutant", variable)) %>%
```

Upon confirmation, the changes will be applied to the full dataset

paper	id	model	variable	method	parameter	type	reason	decision
andersen2008size	1	generalized additive Poisson time series regression model	temperature	smoothing spline	degrees of freedom	parameter	NA	4 or 5 df
andersen2008size	2	generalized additive Poisson time series regression model	dew-point temperature	smoothing spline	degrees of freedom	parameter	NA	4 or 5 df
andersen2008size	3	generalized additive Poisson time series regression model	calendar time	smoothing spline	degrees of freedom	parameter	to control for long-term trend and seasonality	3, 4, or 5 df/year
andersen2008size	4	generalized additive Poisson time series regression model	pollutant	NA	NA	temporal	to include days with the strongest lag effects	4-day pollutant average (lag 0-3
andersen2008size	5	generalized additive Poisson time series regression model	pollutant	NA	NA	temporal	to include days with the strongest lag effects	5-day average (lag 0-4)
andersen2008size	6	generalized additive Poisson time series regression model	pollutant	NA	NA	temporal	to include days with the strongest lag effects	6-day average (lag 0-5)

Fig. 1. The Shiny application interface to validate and standardize Large Language Model (LLM)-generated output. (1) The default interface after loading the input CSV file. (2) The table view will update interactively to reflect the edit: for paper with handle "andersen2008size" and id in 4, 5, 6, modify the variable name *NCtot* to *pollutant*. (3) After clicking the Confirm button, the corresponding `tidyverse` code for the modification is generated, and the table view returns to its original unfiltered view with the edit applied. The edited data can be downloaded by clicking the Download CSV button.

3.5 Summary

Figure 2 summarises the whole workflow proposed for extracting and analyzing decisions from published literature using LLMs. Once researchers have identified a set of literature of interest, a prompt is needed to instruct LLMs to extract decisions from this literature. The outputs from LLM need to be validated and standardized before further analysis, due to the authors' varied writing styles. The validated data can then be used for exploratory data analysis of decisions, and one analysis we propose is to calculate paper similarity. This paper similarity metric can be seen as a distance metric among papers, which can be used for clustering and dimension reduction to visualize the decision patterns among papers.

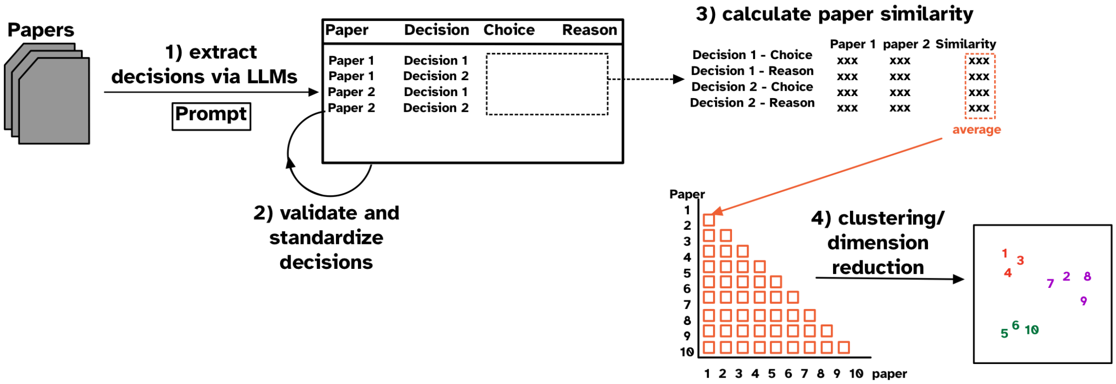

Papers
1) extract decisions via LLMs
Prompt
Paper Decision Choice Reason
Paper 1 Decision 1
Paper 1 Decision 2
Paper 2 Decision 1
Paper 2 Decision 2
2) validate and standardize decisions
3) calculate paper similarity
Paper 1 paper 2 Similarity
Decision 1 - Choice xxx xxx xxx
Decision 1 - Reason xxx xxx xxx
Decision 2 - Choice xxx xxx xxx
Decision 2 - Reason xxx xxx xxx
average
Paper
1 2 3 4 5 6 7 8 9 10
4) clustering/ dimension reduction
1 2 3 4 5 6 7 8 9 10 paper
1 3 4 7 2 8 9 5 6 10

Fig. 2. The workflow for extracting decisions from published literature using Large Language Models (LLMs) and analyzing the extracted decisions. The workflow consists of four main steps: (1) Extract decisions automatically from literature with LLMs, (2) Validate and standardize LLM outputs, (3) Calculate paper similarity and visualization, and (4) visualization with clustering or dimension reduction methods.

4 Results

We apply the workflow to extract the decisions in 56 studies that estimate the effect of particulate matter (PM_{10} and $PM_{2.5}$) on mortality and hospital admission using Gemini (`gemini-2.0-flash`). We focus on the baseline model reported in each paper, excluding secondary models (e.g., lag-distributed models), multi-pollutant models, and alternatives tested in the sensitivity analysis, which are discussed in **?@sec-discussions**. This yields 242 decisions extracted, averaging 4 decisions per paper.

4.1 Validation and standardization of LLM outputs

Table 2. Summary of validation and standardization edits made during the review process.

Reason	Count
Remove decisions out of scope: other pollutants and sensitivity analysis	50
Edit made to recode smoothing parametser unit to per year	45
Duplicates	9

Table 2. Summary of validation and standardization edits made during the review process.

Reason	Count
Fix incorrect capture	9
Edit made due to decisions are too general, e.g. minimum of 1 df per year was required	6
Remove decisions related to definition of variables, e.g. season	5
Total	124

Table 2 summarizes the number of edits made during the review process using the Shiny application. Validation includes fixing incorrect captures, removing non-decision (e.g., definition of variables), removing duplication, excluding irrelevant decisions (e.g., sensitivity analyses), and excluding decisions whose stated reasons reflect general guidelines rather than actual choices (e.g., “minimum of 1 degree of freedom per year is required”).

Standardization is performed on the variable names of decisions and choices. The variable name in the decisions are standardized into four main categories:

- **temperature**: “mean temperature”, “average temperature”, “temperature”, “air temperature”, “ambient temperature”
- **humidity**: “dewpoint temperature” and its hyphenated variants, relative humidity”, “humidity”
- **PM**: “pollutant”, “pollution”, “particulate matter”, “particulate”, “PM10”, “PM2.5”
- **time**: “date”, “time”, “trends”, “trend”

Notice that “dewpoint temperature” is standardized under humidity because it serves as a proxy for temperature in achieving a 100% relative humidity.

Decisions themselves also require standardization. For example, the smoothing parameter (number of knots and degree of freedom) may be expressed as *per year* or *in total*, and temporal lag decision may be expressed in different formats (e.g., “6-day average”, “mean of lags 0+1”, “lagged exposure up to 6 days”). Decision choices on the smoothing parameter are manually recoded to a *per year* basis, as in Table 2. Temporal decisions show a wider variety, which makes manual standardization impractical. However, we observe that they generally fall into two categories:

- **multi-day average lags**: “6-day average”, “3-d moving average”, “mean of lags 0+1”, “cumulative lags, mean 0+1+2”, and
- **single-day lags**: “lagged exposure up to 6 days”, “lag days from 0 to 5”

Hence we apply a secondary LLM (claude-3-7-sonnet-latest) to convert temporal decisions into a consistent format: `multi-day: lag [start]-[end]` and `single-day: lag [start], . . . ,lag [end]`. This converts “6-day average” into “multi-day: lag 0-5” and “lagged exposure up to 6 days” into “single-day: lag 0, lag 1, lag 2, lag 3, lag 4, lag 5”.

4.2 Exploratory analysis of decision choices

As raised in Section 3.1, not all decisions reported in the literature include both the decision choice and the rationale. Some decisions may only report the choice without a stated reason, while others may provide a reason without specifying the exact choice made. Table 3 summarizes the missingness of the decisions and the reason. While 37% of decisions are complete in both decision choices and reasons, 55% of decisions lack a stated rationale for the choice. This reflects a common reporting practice in the field, where authors often report the decision choice used without an explicit reason.

Table 3. Missingness of decision and reason fields in the Gemini-extracted decisions. Most decisions report the choice (35.5 + 57.1 = 92%), but 57.1% lacks a stated reason.

	Decision	
Reason	Non-missing	Missing
Non-missing	90 (37.2%)	14 (5.8%)
Missing	134 (55.4%)	4 (1.7%)

Table 4. Count of variable-type decisions in the Gemini-extracted decisions. The most commonly reported decision are the parameter choices and temporal lags for for time, PM, temperature, and humidity.

Variable	Type	Count
time	parameter	44
PM	temporal	39
temperature	parameter	35
humidity	parameter	25
temperature	temporal	23
humidity	temporal	19
PM	parameter	9
time	temporal	3

Table 4 lists the eight most frequently reported decisions: parameter and temporal choice for `time`, `PM`, `temperature`, and `humidity`. While a wider list of variables has been used in the analysis, these four variables are most commonly included in baseline models. This includes the smoothing parameter used for time, temperature, and humidity in the smoothing method (natural spline and smoothing spline) and temporal lag choices for PM, temperature, and humidity.

Table 5. Options captured for parameter choices for time, humidity, and temperature variables in the Gemini-extracted decisions. The choices for natural spline knots are generally less varied than the degree of freedom choices for smoothing spline. Choices for temperature and humidity tend to be close, given they are both weather related variables, while the choices for time are more varied inherently.

Method	Variable	Decision
natural spline	humidity	3, 4
natural spline	temperature	3, 4, 6
natural spline	time	1, 1.5, 3, 4, 6, 7, 8, 12, 15, 30
smoothing spline	humidity	2, 3, 4, 6, 8, 50% of the data
smoothing spline	temperature	2, 3, 4, 6, 8, 50% of the data
smoothing spline	time	1, 3, 4, 5, 6, 7, 7.7, 8, 9, 10, 12, 30, 100, 5% of the data

Table 5 presents the number of knots or degree of freedom used in two spline methods (natural and smoothing spline) applied to variable `time`, `humidity`,and `temperature`, with all values standardized to a *per year* scale. The choices of knots for natural spline have less variation than the degree of freedom choices for smoothing spline. Choices for temperature and humidity are generally similar, given that they are both weather-related variables, whereas choices for time are more varied. This tabulation provides a reference set for common parameter choices for future studies and helps to identify anomalies and special treatment in practice. For example, the choice of 7.7 degree of freedom reported in Castillejos et al. [12] may prompt analysts to seek further justification for its use. By cross-comparing with other reporting, some decisions appear ambiguous. For example, in Moolgavkar [55] and Moolgavkar [56], the reported value of 30 and 100 degrees of freedom for time may be understandable for experienced domain researchers, but it can be unclear for junior analysts as to whether they refer to the parameter used for the full study period or on a per-year basis, which is often clear in other papers. We also observe a different report style from Schwartz [66], where smoothing spline parameters are expressed as a proportion of the data (“5% of the data” and “5% of the data”), rather than a fixed numerical value.

Table 6. Options captured for temporal lag choices for PM, temperature, and humidity variables in the Gemini-extracted decisions. Both single-day lags and multi-day average lags are commonly used, generally considering up to five days prior (lag 5).

Lag type	Variable	Decision
multi-day average	PM	lag 0-1, 0-2, 0-3, 0-4, 0-5, 0-6
multi-day average	humidity	lag 0-1, 0-2, 0-3, 0-5, 1-5, 2-4
multi-day average	temperature	lag 0-1, 0-2, 0-3, 0-5, 2-4
single-day lag	PM	lag 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13
single-day lag	humidity	lag 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13
single-day lag	temperature	lag 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13

Similarly, Table 6 summarizes the temporal lag choices for `PM`, `temperature`, and `humidity`. For single-day lags, the lags are considered up to 13 days (approximately two weeks) while for multi-day averages, 3-day and 5-day averages are the most common, although other choices such as 2-4 day average are also observed [52].

4.3 Paper similarity calculation, clustering analysis, and visualization

Given the number of decisions reported in Table 4, we focus on the six most common variable-type decisions for calculating paper similarity: parameter choices for time, temperature, and humidity, and temporal lag choices for PM, temperature, and humidity. We also restrict our analysis to papers that report at least three of these six decisions, resulting in 48 papers for the paper similarity calculation. This ensures that the paper similarity metric is based on a sufficient number of comparable decisions. We use the default text embedding model (BERT) in the `text` package and cosine similarity to compute the similarity score. Sensitivity analysis on different text embedding models is checked in Section 4.4.3. Paper similarity is then calculated as the average of decision similarity for each paper pair. The resulting similarity score is then used as the distance matrix in multi-dimensional scaling (MDS) and plotted in Figure 3. The two MDS dimension axes reveal three clusters correspond to the three smoothing methods used in these analyses: LOESS, natural spline, and smoothing spline, where natural spline is commonly used in U.S. based studies suggested in the NMMAPS study [62], while LOESS and smoothing spline are more often used in the European studies, as suggested in the APHEA [39] and APHEA2 [40] project.

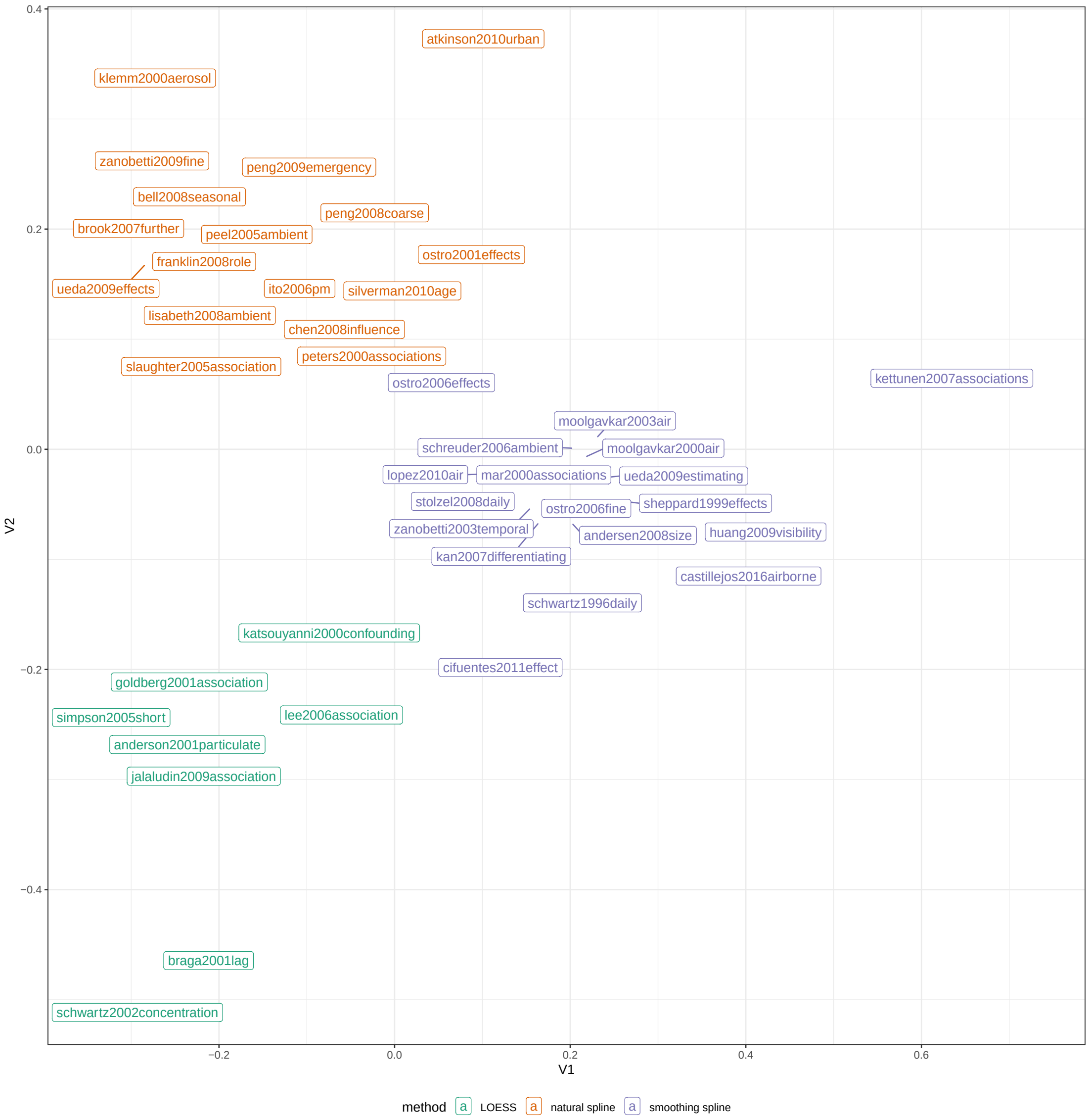


Fig. 3. The multi-dimensional scaling (MDS) based on paper similarity distance for `length(good_pp)` air pollution mortality modeling papers, colored by the smoothing method used. The MDS reveals the three distinct groups of papers, corresponds to LOESS, natural spline, and smoothing spline. These groups corresponds to the different modeling strategies debated in the European and U.S. studies, as documented in the APHENA project [41].

4.4 Sensitivity analysis

A series of sensitivity analysis have been conducted to explore the reproducibility across runs (Section 4.4.1), model providers (Section 4.4.2), and the sensitivity of text model for computing the semantic decision similarity (Section 4.4.3).

4.4.1 LLM reproducibility.

Table 7. Example comparing Gemini's text extraction for Andersen et al. [3] across two runs. The extracted decisions are identical in both runs.

Variable	Run1	Run2
NCtot	6day average (lag 05)	6day average (lag 05)
calendar time	3 4 or 5 dfyear	3 4 or 5 dfyear
dew-point temperature	4 or 5 df	4 or 5 df
temperature	4 or 5 df	4 or 5 df

Table 8. Number of differences in the reason and decision fields across Gemini runs for papers with consistent number of decisions across runs.

Num. of difference	Count	Proportion (%)
0	358	79.73
1	12	2.67
2	8	1.78
3	0	0.00
4	24	5.35
5	12	2.67
6	3	0.67
7	0	0.00
8	10	2.23
9	6	1.34
10	10	2.23
11	6	1.34
Total	449	100.00

We assess the reproducibility across runs of Gemini (`gemini-2.0-flash`) by repeating the text extract task five times for each of the 62 papers and performing pairwise comparison between runs. This generates $5 \times 4/2 \times 62 = 620$ possible comparisons for both "reason" and "decisions" fields. Comparisons are excluded when two runs produced a different number of decisions, since this would require manual alignment. This leaves 449 out of 620 (72%) extractions to compare. Table 7 prints a comparison of decisions in Andersen et al. [3] across two runs, and all four decisions are identical with no difference. Table 8 summarizes the number of differences observed in each pairwise comparison. Among all comparisons, 80% produces the identical text in reason and decision. The discrepancies mainly come from the following two reasons:

1) Gemini extracted the same decision in different lengths. For example, in Kan et al. [37], some runs may extract "singleday lag models underestimate the cumulative effect of pollutants on mortality 2day moving average **of current and previous day concentrations** (lag=01)", while others extract "singleday lag models underestimate the cumulative effect of pollutants on mortality 2day moving average (lag=01)".

2) Gemini fails to extract reasons in some runs but not others. For example, in Burnett et al. [10], the first run generates NA in the reason, but the remaining four runs are identical, with the reason populated. In Ueda et al. [73] and Castillejos et al. [12] , runs 1 and 5 fail to extract the reason and produce the same incomplete version, whereas runs 2, 3, and 4 produce accurate versions with reason populated.

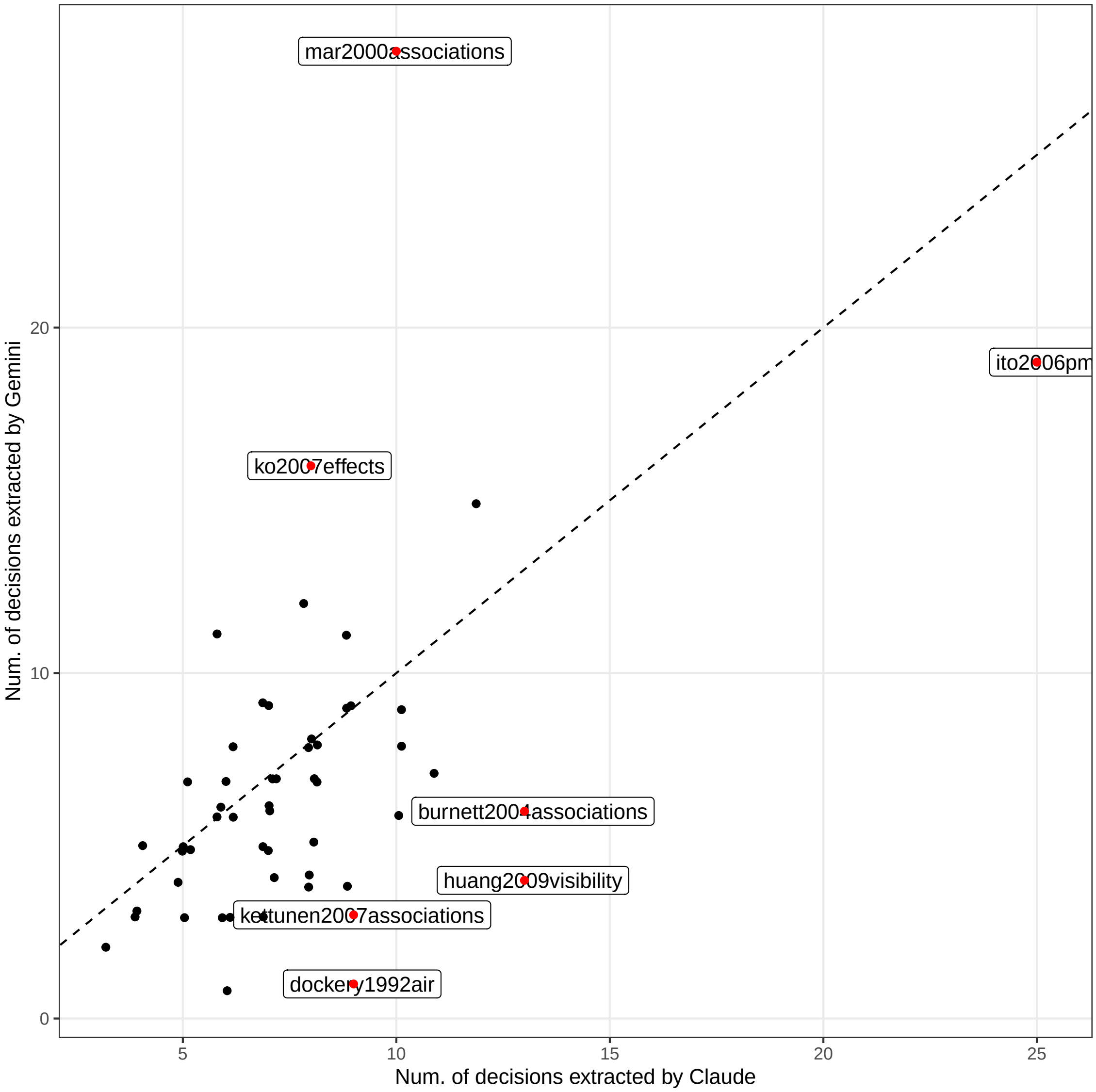


Fig. 4. Comparison of decisions extracted by Claude and Gemini. Each point represents a paper, with the x- and y-axis showing the number of decisions extracted by Claude and Gemini, respectively. The dashed 1:1 line marks where both models extract the same number of decisions. More points fall below this line, suggesting Claude extracts more decisions – often including noise from data pre-processing or secondary data analysis steps – which requires additional manual validation.

4.4.2 LLM models. Reading text from PDF documents require Optical Character Recognition (OCR) to convert images into machine-readable text, which currently is only supported by Antropic Claude and Google Gemini. We compare the number of decisions extracted by Gemini (`gemini-2.0-flash`) and Claude (`claude-3-7-sonnet-latest`) across all 62 papers. In Figure 4, each point represents a paper, with the x- and y-axis showing the number of decisions extracted by Claude and Gemini, respectively. The dashed 1:1 line marks where both models extract the same number of decisions. In general, the two models produce a similar number of decisions. However, more points fall below this line, suggesting Claude extracts more decisions, often including noise from data pre-processing or secondary data analysis steps. Examples of papers with large discrepancies include Mar et al. [53] (Claude: 10 vs. Gemini: 28), Ito et al. [31] (Claude: 25 vs. Gemini: 19), Ko et al. [44] (Claude: 8 vs. Gemini: 16), among others. For both Claude and Gemini, we find they sometimes fail to link the general term "weather variables" to the specific weather variables (e.g., Dockery et al. [17] and Burnett et al. [11] for Gemini and Dockery et al. [17] and Katsouyanni et al. [40] for Claude). Although our prompt specified that some decisions may require linking information across sentences and paragraphs to identify the correct variable, this instruction doesn't appear to be applied consistently.

4.4.3 Text model. We have conducted sensitivity analyses on the text model for calculating the decision similarity score from the Gemini outputs. The tested language models include 1) BERT [16] by Google, 2) RoBERTa [51] by Facebook AI, trained on a larger dataset (160GB v.s. BERT's 15GB), 3) XLNnet [79] by Google Brain, and two domain-trained BERT models: 4) sciBERT [5], trained on scientific literature, and 5) bioBERT [46], trained on PubMed and PMC data.

Figure 5 shows the distribution of the decision similarity and the corresponding multi-dimensional scaling visualization, where distances are calculated from the paper similarity for each text model. At the decision level, the BERT model produces the widest variation across all five models, while the similarity scores from XLNet are all close to 1. While the raw scores are not directly comparable across models due to the difference in the underlying transformer architecture, the visualizations from multi-dimensional scaling (MDS) based on paper similarity scores all show a similar clustering pattern corresponding to the three main smoothing methods (LOESS, natural spline, and smoothing spline).

5 Discussion

5.1 Large-language models for information extraction

Numerous studies [4, 19, 22, 24, 26, 28, 42, 47, 61, 65, 67] have demonstrated the capability of LLMs for information extraction tasks. Our work applies the LLMs to extract analytic decisions in scientific literature, providing further evidence of their effectiveness. Our task requires capturing more complex analytical decisions and their justifications, which typically span more than just a few tokens, like in named entity recognition. Our task also requires linking information across sentences and sometimes sections to correctly identify the variables of a decision (e.g., linking "weather" to "temperature" and "humidity"). While LLM has performed well on extracting decisions from the literature, manual validations are still required to ensure the quality of the extracted decisions for downstream analysis. Most existing applications evaluate LLMs by comparing their outputs to human-annotated datasets, reporting metrics such as precision, recall, and F1 score. Because this approach depends on labeled data, and it is not yet clear how these outputs should be validated for downstream analysis in practice. In our work, we automate some of the manual validation with a secondary LLM (Claude) to standardize the temporal lag choices in different expressions into two categories.

With a default temperature of one and the prompt to instruct the model to extract the original text rather than paraphrase, we find that hallucination is not a major issue with Claude and Gemini in this application. Since LLM outputs are inherently probabilistic, we also conduct sensitivity analyses on reproducibility across runs and model

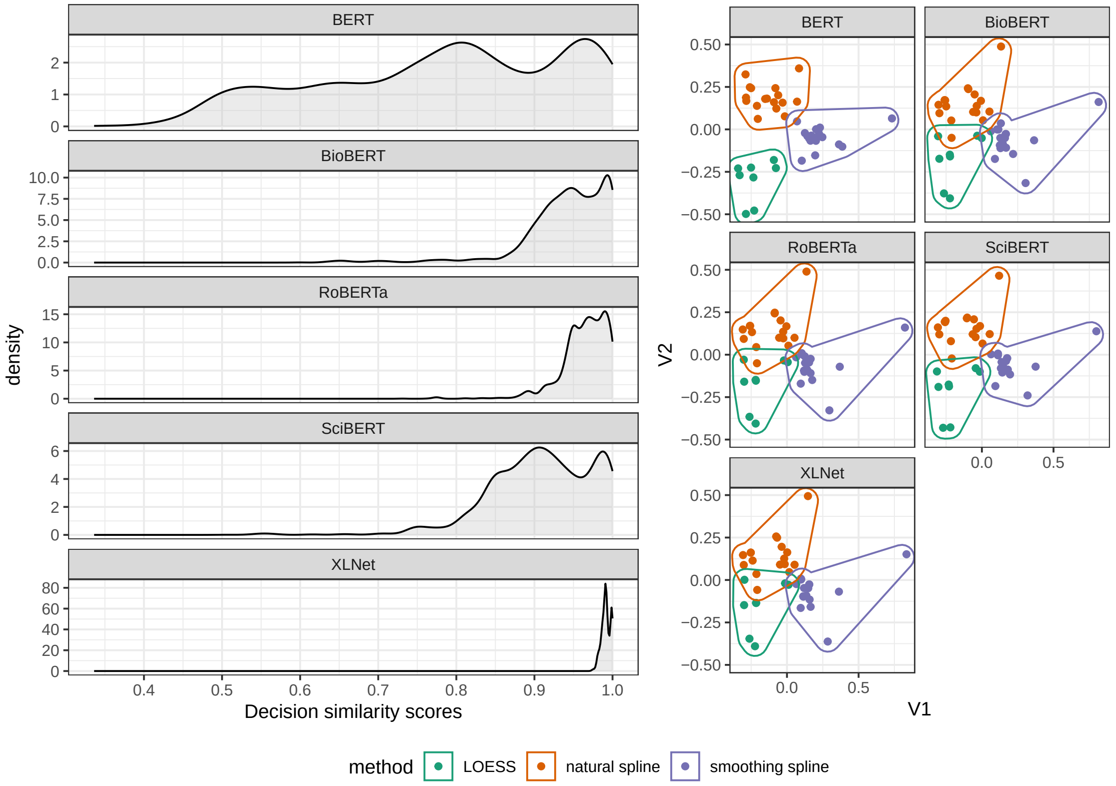


Fig. 5. Distribution of decision similarity (left) and multi-dimensional scaling (MDS) of the paper similarity scores (right) computed for five different text models (BERT, BioBERT, RoBERTa, SciBERT, and XLNet). The default language model, BERT, produces the widest variation across the five models, while the similarity scores form XLNet are all close to 1. The model BioBERT, RoBERTa, and SciBERT yield decision similar scores mostly between 0.7 to 1. All the text models shows a similar clustering structure based on the three main smoothing methods (LOESS, natural spline and smoothing spline).

providers. The output is generally stable: repeated runs with the Gemini produce consistent results, and different models extracted a similar number of decisions.

While we optimize the prompt for decision extraction in this work, an alternative approach is to fine-tune a local model to enhance LLM performance. A catered local model could be useful for extraction decisions for a comprehensive literature review on a larger scale, but it would require greater model training efforts with labeled data.

5.2 Extracting other types of decisions

In this work, we focus on modeling decisions for the baseline model in the air pollution epidemiology literature. Analyses in this field often fit multiple models for different health outcomes and use secondary models, such as distributed lag models and multi-pollutant models, to estimate relative risks and multi-pollutant interactions. These increase the complexity of decision extraction with LLMs because authors often only describe the differences from the baseline specification, implicitly assuming other decisions remain unchanged. Hence, LLMs will need to link the decisions across different models and reconstruct the complete set of decisions for each model.

Beyond modeling choices, decisions in data pre-processing are also interesting to compare. For example, Braga et al. [9] aggregated air pollution measures from multiple PM10 monitors within the same location into a single value. Pre-processing choices such as data source, aggregation method, imputation also have an impact on the uncertainty of the estimated effect size of particulate matter. However, these decisions are often not properly and adequately described in the manuscript, making it impossible to extract by LLMs. Proper documentation and reporting standards in pre-processing decisions are needed before our workflow could be applied to pre-processing decisions.

With growing advocacy for reproducibility, papers nowadays are expected to share code and data, if applicable. Code availability provides a useful supplementary source for identifying decision choices and cross-checking them against descriptions in the manuscript. However, while the script may reveal what choices were made, the rationales behind these choices are often not documented under the current practice.

5.3 Generalizability of the workflow

In principle, our workflow is scalable and generalizable to a random set of applied papers. However, insights about the data analysis practices are more likely to be revealed when papers share certain similarities. For example, literature on the same topic but different authors allows for understanding of common practices within a field, literature using the same methodology across different disciplines allows comparisons of the same statistical method across fields; and literature that considers the same variables can show how those variables are used in different domains.

Our LLM prompt for extracting decisions will need to be customized for each application of the workflow. The general prompt structure and the data schema for recording decisions can be reused, while examples within the prompt may be adapted to suit the specific application. The shiny application for interactively validating and standardizing decisions can be reused across applications. Calculating paper similarity requires comparing decisions on the same variable and type across paper pairs. For papers with limited similarities, the number of comparable decisions may be limited. Diagnostic functions are available to display decisions side by side or provide summary statistics on the number of comparable decisions. Uncertainty visualization on the paper similarity score can be used to highlight the confidence with respect to the number of comparable decisions.

As a new method for collecting analytic decision data from literature, our workflow can be connected to meta-analysis to assess how different decisions influence results. More broadly, it can also be integrated into literature search and recommender systems to suggest similar papers based on the analytic decisions they employ.

6 Conclusion

In this paper, we developed a scalable and generalizable pipeline for automatically extracting analytical decisions using LLMs from scientific literature to study how analysts make decisions in data analysis. We also introduced a method for calculating paper similarity through comparing the similarities among decision choices, and the similarity metric can be used as a distance to cluster papers by their decision choices and visualization with dimension reduction algorithms, such as multidimensional scaling. We applied this pipeline to a set of air pollution modeling literature that associates daily particulate matter and daily mortality and hospital admissions. From the extracted modeling decisions, we identify the most common decision choices in this type of analysis, and the paper similarity score calculation revealed the three clusters of paper corresponding to different smoothing methods.

Many work on studying decision-making in data analysis conduct qualitative interviews with a small number of analysts to understand their decision-making process. “Many-analysts” studies gather together analysts in a controlled experiment to observe analysts conduct the analysis. Our approach is also observational in nature, but we “observe” analysts in real-world problems with real data that have policy implications, while being scalable and cost-effective for a broader exploration of decision-making practices in different contexts and disciplines. Compared to sensitivity analysis or multiverse analysis, our approach offers a different perspective by pooling together decisions made in analyses across the field to reveal the options considered to highlight uncertainty in decisions that require further sensitivity analyses to assess their impact [60, 72].

References

[1] Eric Alexander, Joe Kohlmann, Robin Valenza, Michael Witmore, and Michael Gleicher. 2014 ieee conference on visual analytics science and technology (vast). pages 173–182, 10 2014. doi: 10.1109/VAST.2014.7042493. URL https://ieeexplore.ieee.org/document/7042493.

[2] Sara Alspaugh, Nava Zokaei, Andrea Liu, Cindy Jin, and Marti A. Hearst. Futzing and moseying: Interviews with professional data analysts on exploration practices. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):22–31, 01 2019. doi: 10.1109/TVCG.2018.2865040. URL https://ieeexplore.ieee.org/document/8440815.

[3] Z. J. Andersen, P. Wahlin, O. Raaschou-Nielsen, M. Ketzel, T. Scheike, and S. Loft. Size distribution and total number concentration of ultrafine and accumulation mode particles and hospital admissions in children and the elderly in copenhagen, denmark. *Occupational and Environmental Medicine*, 65(7):458–466, 07 2008. doi: 10.1136/oem.2007.033290. URL https://oem.bmj.com/content/65/7/458. Publisher: BMJ Publishing Group Ltd Section: Original article PMID: 17989204.

[4] Moussa Baddour, Stéphane Paquelet, Paul Rollier, Marie De Tayrac, Olivier Dameron, and Thomas Labbe. 2024 ieee 12th international conference on intelligent systems (is). pages 1–8, 08 2024. doi: 10.1109/IS61756.2024.10705235. URL https://ieeexplore.ieee.org/abstract/document/10705235. ISSN: 2767-9802.

[5] Iz Beltagy, Kyle Lo, and Arman Cohan. Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp). pages 3613–3618, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1371. URL https://www.aclweb.org/anthology/D19-1371.

[6] Steven Bethard and Dan Jurafsky. Cikm ’10: International conference on information and knowledge management. pages 609–618, Toronto ON Canada, 10 2010. ACM. doi: 10.1145/1871437.1871517. URL https://dl.acm.org/doi/10.1145/1871437.1871517.

[7] Graeme Blair, Jasper Cooper, Alexander Coppock, and Macartan Humphreys. Declaring and diagnosing research designs. *American Political Science Review*, 113(3):838–859, 08 2019. doi: 10.1017/S0003055419000194. URL https://www.cambridge.org/core/product/identifier/S0003055419000194/type/journal_article.

[8] Rotem Botvinik-Nezer, Felix Holzmeister, Colin F. Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, Roni Iwanir, Jeanette A. Mumford, R. Alison Adcock, Paolo Avesani, Blazej M. Baczkowski, Aahana Bajracharya, Leah Bakst, Sheryl Ball, Marco Barilari, Nadège Bault, Derek Beaton, Julia Beitner, Roland G. Benoit, Ruud M. W. J. Berkers, Jamil P. Bhanji, Bharat B. Biswal, Sebastian Bobadilla-Suarez, Tiago Bortolini, Katherine L. Bottenhorn, Alexander Bowring, Senne Braem, Hayley R. Brooks, Emily G. Brudner, Cristian B. Calderon, Julia A. Camilleri, Jaime J. Castrellon, Luca Cecchetti, Edna C. Cieslik, Zachary J. Cole, Olivier Collignon, Robert W. Cox, William A. Cunningham, Stefan Czoschke, Kamalaker Dadi, Charles P. Davis, Alberto De Luca, Mauricio R. Delgado, Lysia Demetriou, Jeffrey B. Dennison, Xin Di, Erin W. Dickie, Ekaterina Dobryakova, Claire L. Donnat, Juergen Dukart, Niall W. Duncan, Joke Durnez, Amr Eed, Simon B. Eickhoff, Andrew Erhart, Laura Fontanesi, G. Matthew Fricke, Shiguang Fu, Adriana Galván, Remi Gau, Sarah Genon, Tristan Glatard, Enrico Glerean, Jelle J. Goeman, Sergej A. E. Golowin, Carlos González-García, Krzysztof J. Gorgolewski, Cheryl L. Grady, Mikella A. Green, João F. Guassi Moreira, Olivia Guest, Shabnam Hakimi, J. Paul Hamilton, Roeland Hancock, Giacomo Handjaras, Bronson B. Harry, Colin Hawco, Peer Herholz, Gabrielle Herman, Stephan Heunis, Felix Hoffstaedter, Jeremy Hogeveen, Susan Holmes, Chuan-Peng Hu, Scott A. Huettel, Matthew E. Hughes, Vittorio Iacovella, Alexandru D. Iordan, Peder M. Isager, Ayse I. Isik, Andrew Jahn, Matthew R. Johnson, Tom Johnstone, Michael J. E. Joseph, Anthony C. Juliano, Joseph W. Kable, Michalis Kassinopoulos, Cemal Koba, Xiang-Zhen Kong, Timothy R. Koscik, Nuri Erkut Kucukboyaci, Brice A. Kuhl, Sebastian Kupek, Angela R. Laird, Claus Lamm, Robert Langner, Nina Lauharatanahirun, Hongmi Lee, Sangil Lee, Alexander Leemans, Andrea Leo, Elise Lesage, Flora Li, Monica Y. C. Li, Phui Cheng Lim, Evan N. Lintz, Schuyler W. Liphardt, Annabel B. Losecaat Vermeer, Bradley C. Love, Michael L. Mack, Norberto Malpica, Theo Marins, Camille Maumet, Kelsey McDonald, Joseph T. McGuire, Helena Melero, Adriana S. Méndez Leal, Benjamin Meyer, Kristin N. Meyer, Glad Mihai, Georgios D. Mitsis, Jorge Moll, Dylan M. Nielson, Gustav Nilsonne, Michael P. Notter, Emanuele Olivetti, Adrian I. Onicas, Paolo Papale, Kaustubh R. Patil, Jonathan E. Peelle, Alexandre Pérez, Doris Pischedda, Jean-Baptiste Poline, Yanina Prystauka, Shruti Ray, Patricia A. Reuter-Lorenz, Richard C. Reynolds, Emiliano Ricciardi, Jenny R. Rieck, Anais M. Rodriguez-Thompson, Anthony Romyn, Taylor Salo, Gregory R. Samanez-Larkin, Emilio Sanz-Morales, Margaret L. Schlichting, Douglas H. Schultz, Qiang Shen, Margaret A. Sheridan, Jennifer A. Silvers, Kenny Skagerlund, Alec Smith, David V. Smith, Peter Sokol-Hessner, Simon R. Steinkamp, Sarah M. Tashjian, Bertrand Thirion, John N. Thorp, Gustav Tinghög, Loreen Tisdall, Steven H. Tompson, Claudio Toro-Serey, Juan Jesus Torre Tresols, Leonardo Tozzi, Vuong Truong, Luca Turella, Anna E.

van ‘t Veer, Tom Verguts, Jean M. Vettel, Sagana Vijayarajah, Khoi Vo, Matthew B. Wall, Wouter D. Weeda, Susanne Weis, David J. White, David Wisniewski, Alba Xifra-Porxas, Emily A. Yearling, Sangsuk Yoon, Rui Yuan, Kenneth S. L. Yuen, Lei Zhang, Xu Zhang, Joshua E. Zosky, Thomas E. Nichols, Russell A. Poldrack, and Tom Schonberg. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810): 84–88, 06 2020. doi: 10.1038/s41586-020-2314-9. URL https://www.nature.com/articles/s41586-020-2314-9. Publisher: Nature Publishing Group.

[9] Alfésio Luís Ferreira Braga, Antonella Zanobetti, and Joel Schwartz. The lag structure between particulate air pollution and respiratory and cardiovascular deaths in 10 us cities. *Journal of Occupational and Environmental Medicine*, 43(11):927, 11 2001. URL https://journals.lww.com/joem/fulltext/2001/11000/the_lag_structure_between_particulate_air.1.aspx.

[10] Richard T. Burnett, Sabit Cakmak, Mark E. Raizenne, David Stieb, Renaud Vincent, Daniel Krewski, Jeffrey R. Brook, Owen Philips, and Haluk Ozkaynak. The association between ambient carbon monoxide levels and daily mortality in toronto, canada. *Journal of the Air & Waste Management Association*, 48(8):689–700, 08 1998. doi: 10.1080/10473289.1998.10463718. URL https://www.tandfonline.com/doi/full/10.1080/10473289.1998.10463718.

[11] Richard T. Burnett, Stieb ,Dave , Brook ,Jeffrey R. , Cakmak ,Sabit , Dales ,Robert , Raizenne ,Mark , Vincent ,Renaud , , and Tom Dann. Associations between short-term changes in nitrogen dioxide and mortality in canadian cities. *Archives of Environmental Health: An International Journal*, 59(5):228–236, 05 2004. doi: 10.3200/AEOH.59.5.228-236. URL https://doi.org/10.3200/AEOH.59.5.228-236. Publisher: Routledge _eprint: https://doi.org/10.3200/AEOH.59.5.228-236 PMID: 16201668.

[12] Margarita Castillejos, Borja-Aburto ,Victor H. , Dockery ,Douglas W. , Gold ,Diane R. , , and Dana. Loomis. Airborne coarse particles and mortality. *Inhalation Toxicology*, 12(sup1):61–72, 01 2000. doi: 10.1080/0895-8378.1987.11463182. URL https://doi.org/10.1080/0895-8378.1987.11463182. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/0895-8378.1987.11463182.

[13] Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. Unleashing the potential of prompt engineering for large language models. *Patterns*, 6(6):101260, 06 2025. doi: 10.1016/j.patter.2025.101260. URL https://www.sciencedirect.com/science/article/pii/S2666389925001084.

[14] Chaomei Chen. Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57(3):359–377, 2006. doi: 10.1002/asi.20317. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20317. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.20317.

[15] J. K. Chou and C. K. Yang. Papervis: Literature review made easy. *Computer Graphics Forum*, 30(3):721–730, 2011. doi: 10.1111/j.1467-8659.2011.01921.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8659.2011.01921.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8659.2011.01921.x.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Naacl-hlt 2019. page 4171–4186, Minneapolis, Minnesota, 06 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

[17] Douglas W. Dockery, Joel Schwartz, and John D. Spengler. Air pollution and daily mortality: Associations with particulates and acid aerosols. *Environmental Research*, 59(2):362–373, 12 1992. doi: 10.1016/S0013-9351(05)80042-8. URL https://www.sciencedirect.com/science/article/pii/S0013935105800428.

[18] Marian Dörk, Nathalie Henry Riche, Gonzalo Ramos, and Susan Dumais. Pivotpaths: Strolling through faceted information spaces. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2709–2718, 12 2012. doi: 10.1109/TVCG.2012.252. URL https://ieeexplore.ieee.org/document/6327277.

[19] Saeed Farzi, Soumitra Ghosh, Alberto Lavelli, and Bernardo Magnini. Get the best out of 1b llms: Insights from information extraction on clinical documents. page 266–276, Bangkok, Thailand, 08 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.bionlp-1.21. URL https://aclanthology.org/2024.bionlp-1.21/.

[20] Andrew Gelman and Eric Loken. The statistical crisis in science. *American Scientist*, 102(6):460–465, 12 2014. URL https://www.proquest.com/docview/1616141998/abstract/5E050DCE82414037PQ/1. Num Pages: 6 Place: Research Triangle Park, United States Publisher: Sigma XI-The Scientific Research Society.

[21] Elliot Gould, Hannah S. Fraser, Timothy H. Parker, Shinichi Nakagawa, Simon C. Griffith, Peter A. Vesk, Fiona Fidler, Daniel G. Hamilton, Robin N. Abbey-Lee, Jessica K. Abbott, Luis A. Aguirre, Carles Alcaraz, Irith Aloni, Drew Altschul, Kunal Arekar, Jeff W. Atkins, Joe Atkinson, Christopher M. Baker, Meghan Barrett, Kristian Bell, Suleiman Kehinde Bello, Iván Beltrán, Bernd J. Berauer, Michael Grant Bertram, Peter D. Billman, Charlie K. Blake, Shannon Blake, Louis Bliard, Andrea Bonisoli-Alquati, Timothée Bonnet, Camille Nina Marion Bordes, Aneesh P. H. Bose, Thomas Botterill-James, Melissa Anna Boyd, Sarah A. Boyle, Tom Bradfer-Lawrence, Jennifer Bradham, Jack A. Brand, Martin I. Brengdahl, Martin Bulla, Luc Bussière, Ettore Camerlenghi, Sara E. Campbell, Leonardo L. F. Campos, Anthony Caravaggi, Pedro Cardoso, Charles J. W. Carroll, Therese A. Catanach, Xuan Chen, Heung Ying Janet Chik, Emily Sarah Choy, Alec Philip Christie, Angela Chuang, Amanda J. Chunco, Bethany L. Clark, Andrea Contina, Garth A. Covernton, Murray P. Cox, Kimberly A. Cressman, Marco Crotti, Connor Davidson Crouch, Pietro B. D’Amelio, Alexandra Allison de Sousa, Timm Fabian Döbert, Ralph Dobler, Adam J. Dobson, Tim S. Doherty, Szymon Marian Drobniak, Alexandra Grace Duffy, Alison B. Duncan, Robert P. Dunn, Jamie Dunning, Trishna Dutta, Luke Eberhart-Hertel, Jared Alan Elmore, Mahmoud Medhat Elsherif, Holly M. English, David C. Ensminger, Ulrich Rainer Ernst, Stephen M. Ferguson, Esteban Fernandez-Juricic, Thalita Ferreira-Arruda, John Fieberg, Elizabeth A. Finch, Evan A. Fiorenza, David N. Fisher, Amélie Fontaine, Wolfgang Forstmeier, Yoan Fourcade, Graham S. Frank, Cathryn A. Freund, Eduardo Fuentes-Lillo, Sara L. Gandy, Dustin G. Gannon, Ana I. García-Cervigón, Alexis C. Garretson, Xuezhen Ge, William L. Geary, Charly Géron, Marc Gilles, Antje Girndt, Daniel Gliksman, Harrison B. Goldspiel, Dylan G. E. Gomes, Megan Kate Good, Sarah C. Goslee, J. Stephen Gosnell, Eliza M. Grames, Paolo Gratton, Nicholas M. Grebe, Skye M. Greenler, Maaike Griffioen, Daniel M. Griffith, Frances J. Griffith, Jake J. Grossman, Ali Güncan, Stef Haesen, James G. Hagan, Heather A. Hager, Jonathan Philo Harris, Natasha Dean Harrison, Sarah Syedia Hasnain, Justin Chase Havird, Andrew J. Heaton, María Laura Herrera-Chaustre, Tanner J. Howard, Bin-Yan Hsu, Fabiola Iannarilli, Esperanza C. Iranzo, Erik N. K. Iverson, Saheed Olaide Jimoh,

Douglas H. Johnson, Martin Johnsson, Jesse Jorna, Tommaso Jucker, Martin Jung, Ineta Kačergytė, Oliver Kaltz, Alison Ke, Clint D. Kelly, Katharine Keogan, Friedrich Wolfgang Keppeler, Alexander K. Killion, Dongmin Kim, David P. Kochan, Peter Korsten, Shan Kothari, Jonas Kuppler, Jillian M. Kusch, Malgorzata Lagisz, Kristen Marianne Lalla, Daniel J. Larkin, Courtney L. Larson, Katherine S. Lauck, M. Elise Lauterbur, Alan Law, Don-Jean Léandri-Breton, Jonas J. Lembrechts, Kiara L'Herpiniere, Eva J. P. Lievens, Daniela Oliveira de Lima, Shane Lindsay, Martin Luquet, Ross MacLeod, Kirsty H. Macphie, Kit Magellan, Magdalena M. Mair, Lisa E. Malm, Stefano Mammola, Caitlin P. Mandeville, Michael Manhart, Laura Milena Manrique-Garzon, Elina Mäntylä, Philippe Marchand, Benjamin Michael Marshall, Charles A. Martin, Dominic Andreas Martin, Jake Mitchell Martin, April Robin Martinig, Erin S. McCallum, Mark McCauley, Sabrina M. McNew, Scott J. Meiners, Thomas Merkling, Marcus Michelangeli, Maria Moiron, Bruno Moreira, Jennifer Mortensen, Benjamin Mos, Taofeek Olatunbosun Muraina, Penelope Wrenn Murphy, Luca Nelli, Petri Niemelä, Josh Nightingale, Gustav Nilsonne, Sergio Nolazco, Sabine S. Nooten, Jessie Lanterman Novotny, Agnes Birgitta Olin, Chris L. Organ, Kate L. Ostevik, Facundo Xavier Palacio, Matthieu Paquet, Darren James Parker, David J. Pascall, Valerie J. Pasquarella, John Harold Paterson, Ana Payo-Payo, Karen Marie Pedersen, Grégoire Perez, Kayla I. Perry, Patrice Pottier, Michael J. Proulx, Raphaël Proulx, Jessica L. Pruett, Veronarindra Ramananjato, Finaritra Tolotra Randimbiarison, Onja H. Razafindratsima, Diana J. Rennison, Federico Riva, Sepand Riyahi, Michael James Roast, Felipe Pereira Rocha, Dominique G. Roche, Cristian Román-Palacios, Michael S. Rosenberg, Jessica Ross, Freya E. Rowland, Deusdedith Rugemalila, Avery L. Russell, Suvi Ruuskanen, Patrick Saccone, Asaf Sadeh, Stephen M. Salazar, Kris Sales, Pablo Salmón, Alfredo Sánchez-Tójar, Leticia Pereira Santos, Francesca Santostefano, Hayden T. Schilling, Marcus Schmidt, Tim Schmoll, Adam C. Schneider, Allie E. Schrock, Julia Schroeder, Nicolas Schtickzelle, Nick L. Schultz, Drew A. Scott, Michael Peter Scroggie, Julie Teresa Shapiro, Nitika Sharma, Caroline L. Shearer, Diego Simón, Michael I. Sitvarin, Fabrício Luiz Skupien, Heather Lea Slinn, Grania Polly Smith, Jeremy A. Smith, Rahel Sollmann, Kaitlin Stack Whitney, Shannon Michael Still, Erica F. Stuber, Guy F. Sutton, Ben Swallow, Conor Claverie Taff, Elina Takola, Andrew J. Tanentzap, Rocío Tarjuelo, Richard J. Telford, Christopher J. Thawley, Hugo Thierry, Jacqueline Thomson, Svenja Tidau, Emily M. Tompkins, Claire Marie Tortorelli, Andrew Trlica, Biz R. Turnell, Lara Urban, Stijn Van de Vondel, Jessica Eva Megan van der Wal, Jens Van Eeckhoven, Francis van Oordt, K. Michelle Vanderwel, Mark C. Vanderwel, Karen J. Vanderwolf, Juliana Vélez, Diana Carolina Vergara-Florez, Brian C. Verrelli, Marcus Vinícius Vieira, Nora Villamil, Valerio Vitali, Julien Vollering, Jeffrey Walker, Xanthe J. Walker, Jonathan A. Walter, Pawel Waryszak, Ryan J. Weaver, Ronja E. M. Wedegärtner, Daniel L. Weller, and Shannon Whelan. Same data, different analysts: variation in effect sizes due to analytical decisions in ecology and evolutionary biology. *BMC Biology*, 23(1):35, 02 2025. doi: 10.1186/s12915-024-02101-x. URL https://doi.org/10.1186/s12915-024-02101-x.

[22] Bowen Gu, Vivian Shao, Ziqian Liao, Valentina Carducci, Santiago Romero Brufau, Jie Yang, and Rishi J. Desai. Scalable information extraction from free text electronic health records using large language models. *BMC Medical Research Methodology*, 25(1):23, 01 2025. doi: 10.1186/s12874-025-02470-z. URL https://doi.org/10.1186/s12874-025-02470-z.

[23] Ken Gu, Eunice Jun, and Tim Althoff. Understanding and supporting debugging workflows in multiverse analysis. CHI '23, page 1–19, New York, NY, USA, 04 2023. Association for Computing Machinery. doi: 10.1145/3544548.3581099. URL https://dl.acm.org/doi/10.1145/3544548.3581099.

[24] Sonakshi Gupta, Akhlak Mahmood, Pranav Shetty, Aishat Adeboye, and Rampi Ramprasad. Data extraction from polymer literature using large language models. *Communications Materials*, 5(1):269, 12 2024. doi: 10.1038/s43246-024-00708-9. URL https://www.nature.com/articles/s43246-024-00708-9. Publisher: Nature Publishing Group.

[25] Martin Götz, Abhraneel Sarma, and Ernest H. O'Boyle. The multiverse of universes: A tutorial to plan, execute and interpret multiverses analyses using the r package multiverse. *International Journal of Psychology*, 59(6):1003–1014, 2024. doi: 10.1002/ijop.13229. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ijop.13229. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ijop.13229.

[26] Karlyn K. Harrod, Prabin Bhandari, and Antonios Anastasopoulos. From text to maps: Llm-driven extraction and geotagging of epidemiological data. page 258–270, Miami, Florida, USA, 11 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.nlp4pi-1.24. URL https://aclanthology.org/2024.nlp4pi-1.24/.

[27] Florian Heimerl, Qi Han, Steffen Koch, and Thomas Ertl. Citerivers: Visual analytics of citation patterns. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):190–199, 01 2016. doi: 10.1109/TVCG.2015.2467621. URL https://ieeexplore.ieee.org/document/7192685/authors.

[28] Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, Kirk Roberts, and Hua Xu. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, 31(9):1812–1820, 09 2024. doi: 10.1093/jamia/ocad259. URL https://doi.org/10.1093/jamia/ocad259.

[29] Nick Huntington-Klein, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pralhad Burli, Naibin Chen, Paul Grieco, Godwin Ekpe, Todd Pugatch, Martin Saavedra, and Yaniv Stopnitzky. The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry*, 59(3):944–960, 2021. doi: 10.1111/ecin.12992. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ecin.12992. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecin.12992.

[30] Petra Isenberg, Tobias Isenberg, Michael Sedlmair, Jian Chen, and Torsten Möller. Visualization as seen through its research paper keywords. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):771–780, 01 2017. doi: 10.1109/TVCG.2016.2598827. URL https://ieeexplore.ieee.org/document/7539364.

[31] Kazuhiko Ito, William F. Christensen, Delbert J. Eatough, Ronald C. Henry, Eugene Kim, Francine Laden, Ramona Lall, Timothy V. Larson, Lucas Neas, Philip K. Hopke, and George D. Thurston. Pm source apportionment and health effects: 2. an investigation of intermethod variability in associations between source-apportioned fine particle mass and daily mortality in washington, dc. *Journal of Exposure Science & Environmental Epidemiology*, 16(4):300–310, 07 2006. doi: 10.1038/sj.jea.7500464. URL https://www.nature.com/articles/7500464. Publisher: Nature Publishing Group.

[32] Eunice Jun, Maureen Daum, Jared Roesch, Sarah Chasins, Emery Berger, Rene Just, and Katharina Reinecke. Tea: A high-level language and runtime system for automating statistical analysis. UIST '19, page 591–603, New York, NY, USA, 10 2019. Association for Computing Machinery. doi: 10.1145/3332165.3347940. URL https://dl.acm.org/doi/10.1145/3332165.3347940.

[33] Eunice Jun, Melissa Birchfield, Nicole De Moura, Jeffrey Heer, and René Just. Hypothesis formalization: Empirical findings, software limitations, and design implications. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 29(1):1–28, 2022.

[34] Eunice Jun, Audrey Seo, Jeffrey Heer, and René Just. Tisane: Authoring statistical models via formal reasoning from conceptual and data relationships. CHI '22, page 1–16, New York, NY, USA, 04 2022. Association for Computing Machinery. doi: 10.1145/3491102.3501888. URL https://dl.acm.org/doi/10.1145/3491102.3501888.

[35] Alex Kale, Matthew Kay, and Jessica Hullman. Decision-making under uncertainty in research synthesis: Designing for the garden of forking paths. CHI '19, page 1–14, New York, NY, USA, 05 2019. Association for Computing Machinery. doi: 10.1145/3290605.3300432. URL https://dl.acm.org/doi/10.1145/3290605.3300432.

[36] Alex Kale, Sarah Lee, Terrance Goan, Elizabeth Tipton, and Jessica Hullman. Metaexplorer : Facilitating reasoning with epistemic uncertainty in meta-analysis. CHI '23, page 1–14, New York, NY, USA, 04 2023. Association for Computing Machinery. doi: 10.1145/3544548.3580869. URL https://dl.acm.org/doi/10.1145/3544548.3580869.

[37] Haidong Kan, Stephanie J. London, Guohai Chen, Yunhui Zhang, Guixiang Song, Naiqing Zhao, Lili Jiang, and Bingheng Chen. Differentiating the effects of fine and coarse particles on daily mortality in shanghai, china. *Environment International*, 33(3):376–384, 04 2007. doi: 10.1016/j.envint.2006.12.001. URL https://www.sciencedirect.com/science/article/pii/S0160412006002108.

[38] Sean Kandel, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer. Enterprise data analysis and visualization: An interview study. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2917–2926, 12 2012. doi: 10.1109/TVCG.2012.219. URL https://ieeexplore.ieee.org/document/6327298.

[39] K Katsouyanni, J Schwartz, C Spix, G Touloumi, D Zmirou, A Zanobetti, B Wojtyniak, J M Vonk, A Tobias, A Pönkä, S Medina, L Bachárová, and H R Anderson. Short term effects of air pollution on health: a european approach using epidemiologic time series data: the aphea protocol. *Journal of Epidemiology and Community Health*, 50(Suppl 1):S12–S18, 04 1996. doi: 10.1136/jech.50.suppl_1.s12. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1060882/. PMID: 8758218 PMCID: PMC1060882.

[40] Klea Katsouyanni, Giota Touloumi, Evangelia Samoli, Alexandros Gryparis, Alain Le Tertre, Yannis Monopolis, Giuseppe Rossi, Denis Zmirou, Ferran Ballester, Azedine Boumghar, Hugh Ross Anderson, Bogdan Wojtyniak, Anna Paldy, Rony Braunstein, Juha Pekkanen, Christian Schindler, and Joel Schwartz. Confounding and effect modification in the short-term effects of ambient particles on total mortality: Results from 29 european cities within the aphea2 project. *Epidemiology*, 12(5):521, 09 2001. URL https://journals.lww.com/epidem/fulltext/2001/09000/confounding_and_effect_modification_in_the.11.aspx.

[41] Klea Katsouyanni, Jonathan M. Samet, H. Ross Anderson, Richard Atkinson, Alain Le Tertre, Sylvia Medina, Evangelia Samoli, Giota Touloumi, Richard T. Burnett, Daniel Krewski, Tim Ramsay, Francesca Dominici, Roger D. Peng, Joel Schwartz, and Antonella Zanobetti. Air pollution and health: A european and north american approach (aphena). Research Report 142, Health Effects Institute, Boston, MA, 2009.

[42] Uri Katz, Mosh Levy, and Yoav Goldberg. Findings of the association for computational linguistics: Emnlp 2024. pages 8838–8855, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.516. URL https://aclanthology.org/2024.findings-emnlp.516.

[43] Oscar Kjell, Salvatore Giorgi, and H. Andrew Schwartz. The text-package: An r-package for analyzing and visualizing human language using natural language processing and deep learning. *Psychological Methods*, 2023. doi: 10.1037/met0000542. URL https://pubmed.ncbi.nlm.nih.gov/37126041/.

[44] F. W. S. Ko, W. Tam, T. W. Wong, C. K. W. Lai, G. W. K. Wong, T.-F. Leung, S. S. S. Ng, and D. S. C. Hui. Effects of air pollution on asthma hospitalization rates in different age groups in hong kong. *Clinical & Experimental Allergy*, 37(9):1312–1319, 2007. doi: 10.1111/j.1365-2222.2007.02791.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2222.2007.02791.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2222.2007.02791.x.

[45] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

[46] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 02 2020. doi: 10.1093/bioinformatics/btz682. URL https://academic.oup.com/bioinformatics/article/36/4/1234/5566506.

[47] Ni Li, Shorouq Zahra, Mariana Brito, Clare Flynn, Olof Görnerup, Koffi Worou, Murathan Kurfali, Chanjuan Meng, Wim Thiery, Jakob Zscheischler, Gabriele Messori, and Joakim Nivre. Proceedings of the 1st workshop on natural language processing meets climate change (climatenlp 2024). pages 93–110, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.climatenlp-1.7. URL https://aclanthology.org/2024.climatenlp-1.7.

[48] Jiali Liu, Nadia Boukhelifa, and James R. Eagan. Understanding the Role of Alternatives in Data Analysis Practices. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):66–76, January 2020. ISSN 1941-0506. doi: 10.1109/TVCG.2019.2934593. URL https://ieeexplore.ieee.org/document/8805460/.

[49] Yang Liu, Tim Althoff, and Jeffrey Heer. Paths explored, paths omitted, paths obscured: Decision points & selective reporting in end-to-end data analysis. CHI '20, page 1–14, New York, NY, USA, 04 2020. Association for Computing Machinery. doi: 10.1145/3313831.3376533. URL https://dl.acm.org/doi/10.1145/3313831.3376533.

[50] Yang Liu, Alex Kale, Tim Althoff, and Jeffrey Heer. Boba: Authoring and visualizing multiverse analyses. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1753–1763, 02 2021. doi: 10.1109/TVCG.2020.3028985. URL https://ieeexplore.ieee.org/document/9216579/.

[51] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. doi: 10.48550/arXiv.1907.11692.

[52] Elena López-Villarrubia, Ferran Ballester, Carmen Iñiguez, and Nieves Peral. Air pollution and mortality in the canary islands: a time-series analysis. *Environmental Health*, 9:8, 02 2010. doi: 10.1186/1476-069X-9-8. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2843667/. PMID: 20152037 PMCID: PMC2843667.

[53] T F Mar, G A Norris, J Q Koenig, and T V Larson. Associations between air pollution and mortality in phoenix, 1995-1997. *Environmental Health Perspectives*, 108(4):347–353, 04 2000. doi: 10.1289/ehp.00108347. URL https://ehp.niehs.nih.gov/doi/abs/10.1289/ehp.00108347. Publisher: Environmental Health Perspectives.

[54] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. volume 26. Curran Associates, Inc., 2013. URL https://papers.nips.cc/paper_files/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.

[55] Suresh H. Moolgavkar. Air pollution and hospital admissions for diseases of the circulatory system in three u.s. metropolitan areas. *Journal of the Air & Waste Management Association*, 50(7):1199–1206, 07 2000. doi: 10.1080/10473289.2000.10464162. URL https://doi.org/10.1080/10473289.2000.10464162. Publisher: Taylor & Francis.

[56] Suresh H. Moolgavkar. Air pollution and daily mortality in two u.s. counties: Season-specific analyses and exposure-response relationships. *Inhalation Toxicology*, 15(9):877–907, 01 2003. doi: 10.1080/08958370390215767. URL https://doi.org/10.1080/08958370390215767. Publisher: Taylor & Francis.

[57] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticæ Investigationes*, 30(1):3–26, 01 2007. doi: 10.1075/li.30.1.03nad. URL https://www.jbe-platform.com/content/journals/10.1075/li.30.1.03nad. Publisher: John Benjamins.

[58] Arpit Narechania, Alireza Karduni, Ryan Wesslen, and Emily Wall. Vitality: Promoting serendipitous discovery of academic literature with transformers & visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):486–496, 01 2022. doi: 10.1109/TVCG.2021.3114820. URL https://ieeexplore.ieee.org/document/9552447/.

[59] Bart Ostro, Rachel Broadwin, Shelley Green, Wen-Ying Feng, and Michael Lipsett. Fine particulate air pollution and mortality in nine california counties: Results from calfine. *Environmental Health Perspectives*, 114(1):29–33, 01 2006. doi: 10.1289/ehp.8335. URL https://ehp.niehs.nih.gov/doi/10.1289/ehp.8335. Publisher: Environmental Health Perspectives.

[60] Roger D. Peng, Francesca Dominici, and Thomas A. Louis. Model choice in time series studies of air pollution and mortality. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 169(2):179–203, 03 2006. doi: 10.1111/j.1467-985X.2006.00410.x. URL https://doi.org/10.1111/j.1467-985X.2006.00410.x.

[61] Maciej P. Polak and Dane Morgan. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications*, 15(1):1569, 02 2024. doi: 10.1038/s41467-024-45914-8. URL https://www.nature.com/articles/s41467-024-45914-8. Publisher: Nature Publishing Group.

[62] Jonathan M. Samet, Francesca Dominici, Frank C. Curriero, Ivan Coursac, and Scott L. Zeger. Fine particulate air pollution and mortality in 20 u.s. cities, 1987–1994. *New England Journal of Medicine*, 343(24):1742–1749, 12 2000. doi: 10.1056/NEJM200012143432401. URL https://www.nejm.org/doi/full/10.1056/NEJM200012143432401. Publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJM200012143432401.

[63] Abhraneel Sarma, Alex Kale, Michael Moon, Nathan Taback, Fanny Chevalier, Jessica Hullman, and Matthew Kay. multiverse: Multiplexing alternative data analyses in r notebooks (version 0.6.2). *OSF Preprints*, 2021. URL https://github.com/MUCollective/multiverse.

[64] Marko Sarstedt, Susanne J. Adler, Christian M. Ringle, Gyeongcheol Cho, Adamantios Diamantopoulos, Heungsun Hwang, and Benjamin D. Liengaard. Same model, same data, but different outcomes: Evaluating the impact of method choices in structural equation modeling. *Journal of Product Innovation Management*, 41(6):1100–1117, 2024. doi: 10.1111/jpim.12738. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jpim.12738. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jpim.12738.

[65] Mara Schilling-Wilhelmi, Martiño Ríos-García, Sherjeel Shabih, María Victoria Gil, Santiago Miret, Christoph T. Koch, José A. Márquez, and Kevin Maik Jablonka. From text to insight: large language models for chemical data extraction. *Chemical Society Reviews*, 54(3):1125–1150, 2025. doi: 10.1039/D4CS00913D. URL https://pubs.rsc.org/en/content/articlelanding/2025/cs/d4cs00913d. Publisher: Royal Society of Chemistry.

[66] Joel Schwartz. The distributed lag between air pollution and daily deaths. *Epidemiology*, 11(3):320–326, 2000. URL https://www.jstor.org/stable/3703220. Publisher: Lippincott Williams & Wilkins.

[67] Veronica Sciannameo, Daniele Jahier Pagliari, Sara Urru, Piercesare Grimaldi, Honoria Ocagli, Sara Ahsani-Nasab, Rosanna Irene Comoretto, Dario Gregori, and Paola Berchialla. Information extraction from medical case reports using openai instructgpt. *Computer Methods and Programs in Biomedicine*, 255:108326, 10 2024. doi: 10.1016/j.cmpb.2024.108326. URL https://www.sciencedirect.com/science/article/pii/S0169260724003195.

[68] R. Silberzahn, E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtrey, Š. Bahník, F. Bai, C. Bannard, E. Bonnier, R. Carlsson, F. Cheung, G. Christensen, R. Clay, M. A. Craig, A. Dalla Rosa, L. Dam, M. H. Evans, I. Flores Cervantes, N. Fong, M. Gamez-Djokic, A. Glenz, S. Gordon-McKeon, T. J. Heaton, K. Hederos, M. Heene, A. J. Hofelich Mohr, F. Högden, K. Hui, M. Johannesson, J. Kalodimos, E. Kaszubowski, D. M. Kennedy, R. Lei, T. A. Lindsay, S. Liverani, C. R. Madan, D. Molden, E. Molleman, R. D. Morey, L. B. Mulder, B. R. Nijstad, N. G. Pope, B. Pope, J. M. Prenoveau, F. Rink, E. Robusto, H. Roderique, A. Sandberg, E. Schlüter, F. D. Schönbrodt, M. F. Sherman, S. A. Sommer, K. Sotak, S. Spain, C. Spörlein, T. Stafford, L. Stefanutti, S. Tauber, J. Ullrich, M. Vianello, E.-J. Wagenmakers, M. Witkowiak, S. Yoon, and B. A. Nosek. Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, 1(3):337–356, 09 2018. doi: 10.1177/2515245917747646. URL https://doi.org/10.1177/2515245917747646. Publisher: SAGE Publications Inc.

[69] Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366, 11 2011. doi: 10.1177/0956797611417632. URL https://doi.org/10.1177/0956797611417632. Publisher: SAGE Publications Inc.

[70] Jan Simson, Fiona Draxler, Samuel Mehr, and Christoph Kern. Preventing harmful data practices by using participatory input to navigate the machine learning multiverse. CHI '25, page 1–30, New York, NY, USA, 04 2025. Association for Computing Machinery. doi: 10.1145/3706598.3713482. URL https://dl.acm.org/doi/10.1145/3706598.3713482.

[71] Imad Tbahriti, Christine Chichester, Frédérique Lisacek, and Patrick Ruch. Using argumentation to retrieve articles with similar citations: An inquiry into improving related articles search in the medline digital library. *International Journal of Medical Informatics*, 75(6):488–495, 06 2006. doi: 10.1016/j.ijmedinf.2005.06.007. URL https://www.sciencedirect.com/science/article/pii/S1386505605000894.

[72] G. Touloumi, E. Samoli, M. Pipikou, A. Le Tertre, R. Atkinson, and K. Katsouyanni. Seasonal confounding in air pollution and health time-series studies: effect on air pollution effect estimates. *Statistics in Medicine*, 25(24):4164–4178, 2006. doi: 10.1002/sim.2681. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2681. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.2681.

[73] Kayo Ueda, Nitta ,Hiroshi , Ono ,Masaji , , and Ayano Takeuchi. Estimating mortality effects of fine particulate matter in japan: A comparison of time-series and case-crossover analyses. *Journal of the Air & Waste Management Association*, 59(10):1212–1218, 10 2009. doi: 10.3155/1047-3289.59.10.1212. URL https://doi.org/10.3155/1047-3289.59.10.1212. Publisher: Taylor & Francis _eprint: https://doi.org/10.3155/1047-3289.59.10.1212.

[74] Jelte M. Wicherts, Coosje L. S. Veldkamp, Hilde E. M. Augusteijn, Marjan Bakker, Robbie C. M. van Aert, and Marcel A. L. M. van Assen. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, 7, 11 2016. doi: 10.3389/fpsyg.2016.01832. URL https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2016.01832/full. Publisher: Frontiers.

[75] Hadley Wickham. Tidy data. *Journal of Statistical Software*, 59:1–23, 09 2014. doi: 10.18637/jss.v059.i10. URL https://doi.org/10.18637/jss.v059.i10.

[76] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.

[77] Hadley Wickham, Joe Cheng, and Aaron Jacobs. *ellmer: Chat with Large Language Models*, 2025. URL https://CRAN.R-project.org/package=ellmer. R package version 0.1.1.

[78] Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. Large language models for generative information extraction: A survey. doi: 10.48550/arXiv.2312.17617.

[79] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. doi: 10.48550/arXiv.1906.08237.