An LLM-based Pipeline for Understanding Decision Choices in Data Analysis from Published Literature

H. Sherry Zhang[1], Roger D. Peng[1]

ARTICLE HISTORY
Compiled February 1, 2026

[1] University of Texas at Austin, Austin, USA

ABSTRACT
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.

KEYWORDS
decision choice; data analysis; Large Language Models

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

CONTACT: H. Sherry Zhang. Email: hsherryzhang@utexas.edu.

dataset to answer the same research question (Silberzahn et al. 2018; Botvinik-Nezer et al. 2020; Gould et al. 2025). This variation in analytical decision-making, described by Gelman and Loken (2014) 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 (Sarma et al. 2021; Blair et al. 2019). However, what one 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 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) (Harrod et al. 2024; Katz et al. 2024; Farzi et al. 2024; Hu et al. 2024; Sciannameo et al. 2024; Gu et al. 2025; Schilling-Wilhelmi et al. 2025; Gupta et al. 2024; Li et al. 2024; Baddour et al. 2024; Polak and Morgan 2024).

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 the smoothing methods used: LOESS, natural spline, and smoothing spline. These findings align with the APHENA project (Katsouyanni et al. 2009), which synthesizes research from multiple studies in Europe and North America. 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 (Jun, Seo, et al. 2022; Jun, Birchfield, et al. 2022; Jun et al. 2019) that involves 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 (2014) 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. (2018) asks 29 teams 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, in structural equation modeling (Sarstedt et al. 2024), applied microeconomics (Huntington-Klein et al. 2021), neuroimaging (Botvinik-Nezer et al. 2020), and ecology and evolutionary biology (Gould et al. 2025).

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 (Katsouyanni et al. 2009) and domain applications, e.g., fairness machine learning (Simson et al. 2025). Through experiments (Wicherts et al. 2016; Simmons et al. 2011), research has shown that analysts' decisions can lead to p-hacking and inflated effect size when not properly used. 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 (Kale et al. 2019; Y. Liu et al. 2020), interviews with data analysts about exploratory data analysis practice in industry (Alspaugh et al. 2019; Kandel et al. 2012), and about how they consider alternatives in data analysis (J. Liu et al. 2020).

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` (Jun et al. 2019), which supports general statistical analysis; `Tisane` (Jun, Seo, et al. 2022), which guides choices in generalized linear mixed-effects models (GLMMs); and `MetaExplore` (Kale et al. 2023), which accounts for epistemic uncertainty (decision uncertainty) in meta-analysis. The `DeclareDesign` package (Blair et al. 2019) 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 (Sarma et al. 2021; Götz et al. 2024) and visualization of multiverse results (Liu et al. 2021), and debugging tools (Gu et al. 2023).

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 (Lafferty et al., n.d.), word embeddings such as word2vec (Mikolov et al. 2013), and transformer-based architectures like BERT (Devlin et al. 2019), 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 (Harrod et al. 2024), scientific literature (Katz et al. 2024), clinical data (Farzi et al. 2024; Hu et al. 2024; Sciannameo et al. 2024; Gu et al. 2025), chemistry knowledge (Schilling-Wilhelmi et al. 2025), and polymer science (Gupta et al. 2024), climate extreme impact (Li et al. 2024), phenotypes (Baddour et al. 2024), and material properties (Polak and Morgan 2024). 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

(Nadeau and Sekine 2007). 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 (Isenberg et al. 2017), citation information (Chen 2006), e.g., citation list and co-citation, or combinations with other relevant paper metadata (Bethard and Jurafsky 2010; Chou and Yang 2011; Dörk et al. 2012; Heimerl et al. 2016), e.g., author and title. Recent approaches incorporate text-based information using topic modeling (Alexander et al. 2014), argumentation-based information retrieval (Tbahriti et al. 2006), and text embedding (Narechania et al. 2022). 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*

To analyze decisions, we first need to translate free-text descriptions of decisions in academic papers into a tabular format. We record decision following the tidy data principle (Wickham 2014), which states that each variable forms a column and each observation forms a row. For our purpose, each row represents a single decision made in a paper, and an analysis typically involves multiple decisions.

A decision generally consists of three components: the context, the reason, and the decision itself. To characterize the context of a decision, additional information is often required, such as what variable the decision is acted on, what statistical method the decision uses, and what parameter of the method is being decided. Some decisions do not involve a specific method or parameter, for example, a model may be estimated separately for each city rather than jointly across all cities. To account for this, we introduce a type identifier to distinguish whether a decision is parameter-, spatial-, or temporal-based. Method and parameter specifications are not required for spatial or temporal decisions at the model level. The resulting structure for recording decisions is as follows:

- `type`: one of “parameter”, “spatial”, or “temporal”.
- `variable`: the variable to which the statistical method is applied.
- `method`: the statistical method used, (e.g. “LOESS”, “smoothing spline”, “natural spline”).
- `parameter`: the parameter of the method being decided, (e.g. “degrees of freedom”, “number of knots”).
- `reason`: the justification for the decision.
- `decision`: the final choice made.
- `reference`: any cited sources supporting the decision.

For the ease of tracking, we extract the original text describing each decision verbatim, without paraphrasing or summary.

3.2. *Extract decisions automatically from literature with LLMs*

Manually extracting decisions from published papers is labor-intensive and time-consuming. Large language models (LLMs) make it possible to automate this process by providing a collection of PDF documents along with a structured prompt. In the prompt, the LLM is assigned the role of an applied statistician and instructed to extract decisions from each PDF document according to the format described in Section 3.1. The prompt also notes that the reason and decision fields may be missing, with examples provided. We use the `ellmer` package (Wickham et al. 2025) in R to interface with Anthropic Claude and Google Gemini for this task and the full prompt is provided in the Appendix.

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 that semantically equivalent terms are represented in a consistent form. 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 for interactively reviewing and editing the LLM outputs. The Shiny application takes a CSV file as the input and allows users to perform three types of edits: 1) *overwrite*: modify the content of a particular cell, 2) *delete*: remove an irrelevant row (decision), and 3) *add*: manually enter a row (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 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` (Wickham et al. 2019) 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

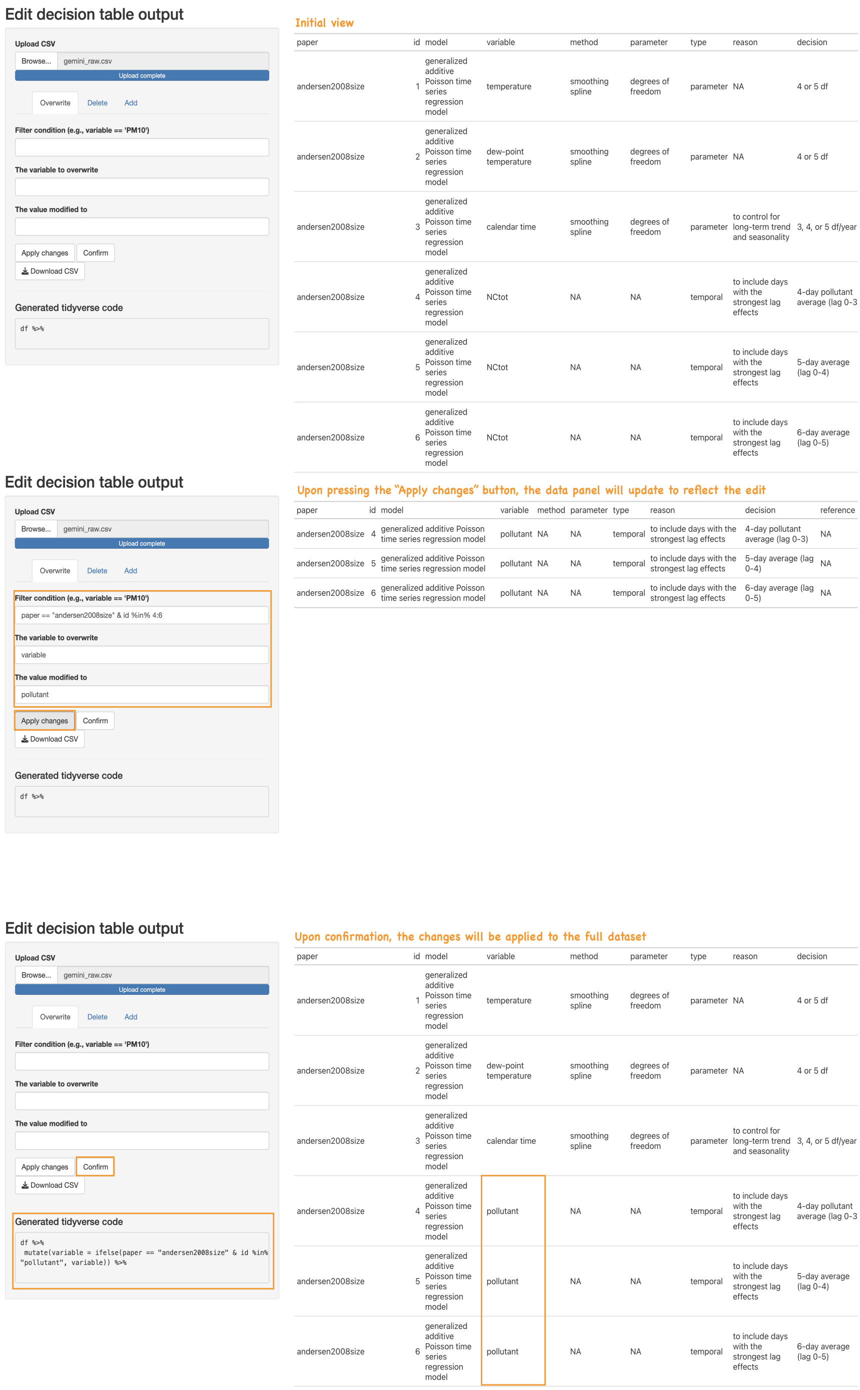

Figure 1. The Shiny application interface to validate and standardize Large Language Model (LLM)-generated output. (1) The default interface after loading the input CSV file. (2) The table view will update interactively to reflect the edit: for paper with handle "andersen2008size" and id in 4, 5, 6, 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.

similarity measure to compare how similar decisions are between paper pairs. A decision is considered comparable between two papers if they share the same variable and decision type, for example, a parameter decision on temperature. To quantify similarity between matched decisions, we consider three aspects: 1) similarity of the decision choice, 2) similarity of the stated reasons, and 3) for parameter decisions, similarity of the statistical method used. Similarity in choice and method reflects the same analytical decision, while similarity in reason reflects a shared rationale, even when the choices differ due to differences in data.

To measure similarity in choices and reasons, we obtain text embeddings and compute cosine similarity using the BERT model via the `text` package (Kjell et al. 2023) in `R`. Method similarity is encoded as a binary indicator (1 if the two papers used the same method, and 0 otherwise) since semantic similarity in text cannot reliably distinguish statistical methods. For example, the textual difference between "smoothing spline" and "natural spline" does not fully capture their methodological distinction. The overall paper similarity is computed as the average similarity across all 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. 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.

3.5. *Summary*

Figure 2 summarises the workflow proposed for extracting and analyzing decisions from published literature using LLMs. After identifying a set of relevant papers, a prompt has been designed to guide the LLM in extracting decisions from the documents. The extracted outputs are then validated and standardized before further analysis. The resulting dataset can then be used for exploratory data analysis and the construction of a paper similarity metric. This measure can be interpreted as a distance metric between papers and used for clustering and dimension reduction to visualize patterns in decision-making across the literature.

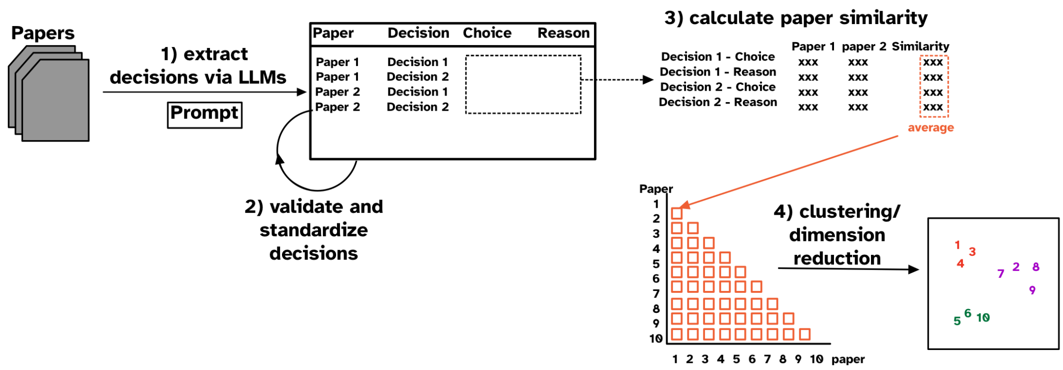

Figure 2. The workflow for extracting decisions from published literature using Large Language Models (LLMs) and analyzing the extracted decisions. The workflow consists of four main steps: (1) Extract decisions automatically from literature with LLMs, (2) Validate and standardize LLM outputs, (3) Calculate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

4. Application

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.

We apply the workflow to extract the decisions in 56 studies reviewed in Atkinson et al. (2014) 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 Section 5. This yields 242 decisions extracted, averaging 4 decisions per paper.

4.1. *Validation and standardization of LLM outputs*

Table 1. 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
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

Table 1. Summary of validation and standardization edits made during the review process.

Reason	Count
Total	124

Table 1 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",

Table 2. 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%)

"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 1. 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*

In practice, data analysis decisions in academic papers are generally not presented individually in the format described in Section 3.1. Authors may combine multiple related decisions into a single sentence for brevity, or omit certain components, not providing a reason for a decision or not stating the exact choice made. Table 2 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. 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 3 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 4. 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 4 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. (2000) may prompt analysts to seek further justification for its use. By cross-comparing with other reporting, some decisions appear ambiguous. For example, in Moolgavkar (2000) and Moolgavkar (2003), 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 (2000), 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 5. 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 5 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 (López-Villarrubia et al. 2010).

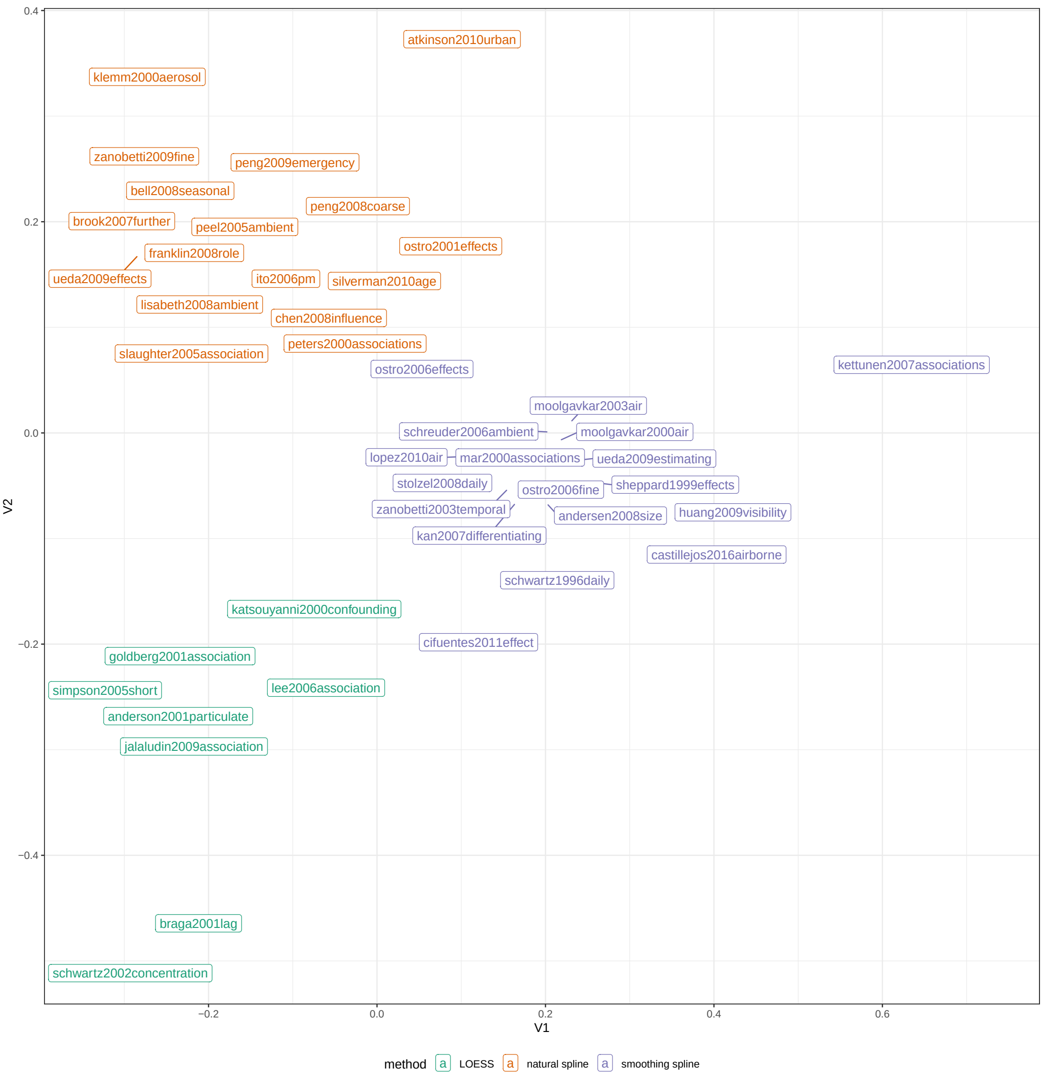

Figure 3. The multi-dimensional scaling (MDS) based on paper similarity distance for `length(good_pp)` air pollution mortality modeling papers, colored by the smoothing method used. The MDS reveals the three distinct groups of papers, corresponds to LOESS, natural spline, and smoothing spline. These groups corresponds to the different modeling strategies debated in the European and U.S. studies, as documented in the APHENA project (Katsouyanni et al. 2009).

4.3. ***Paper similarity calculation, clustering analysis, and visualization***

Given the number of decisions reported in Table 3, 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 (Samet et al. 2000), while LOESS and smoothing spline are more often used in the European studies, as suggested in the APHEA (Katsouyanni et al. 1996) and APHEA2 (Katsouyanni et al. 2001) project.

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 6. Example comparing Gemini's text extraction for Andersen et al. (2008) 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 7. 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 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 6 prints a comparison of decisions in Andersen et al. (2008) across two runs, and all four decisions are identical with no difference. Table 7 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. (2007), 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. (1998), the first run generates `NA` in the reason, but the remaining four runs are identical, with the reason populated. In Ueda et al. (2009) and Castillejos et al. (2000) , 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.

4.4.2. LLM models

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

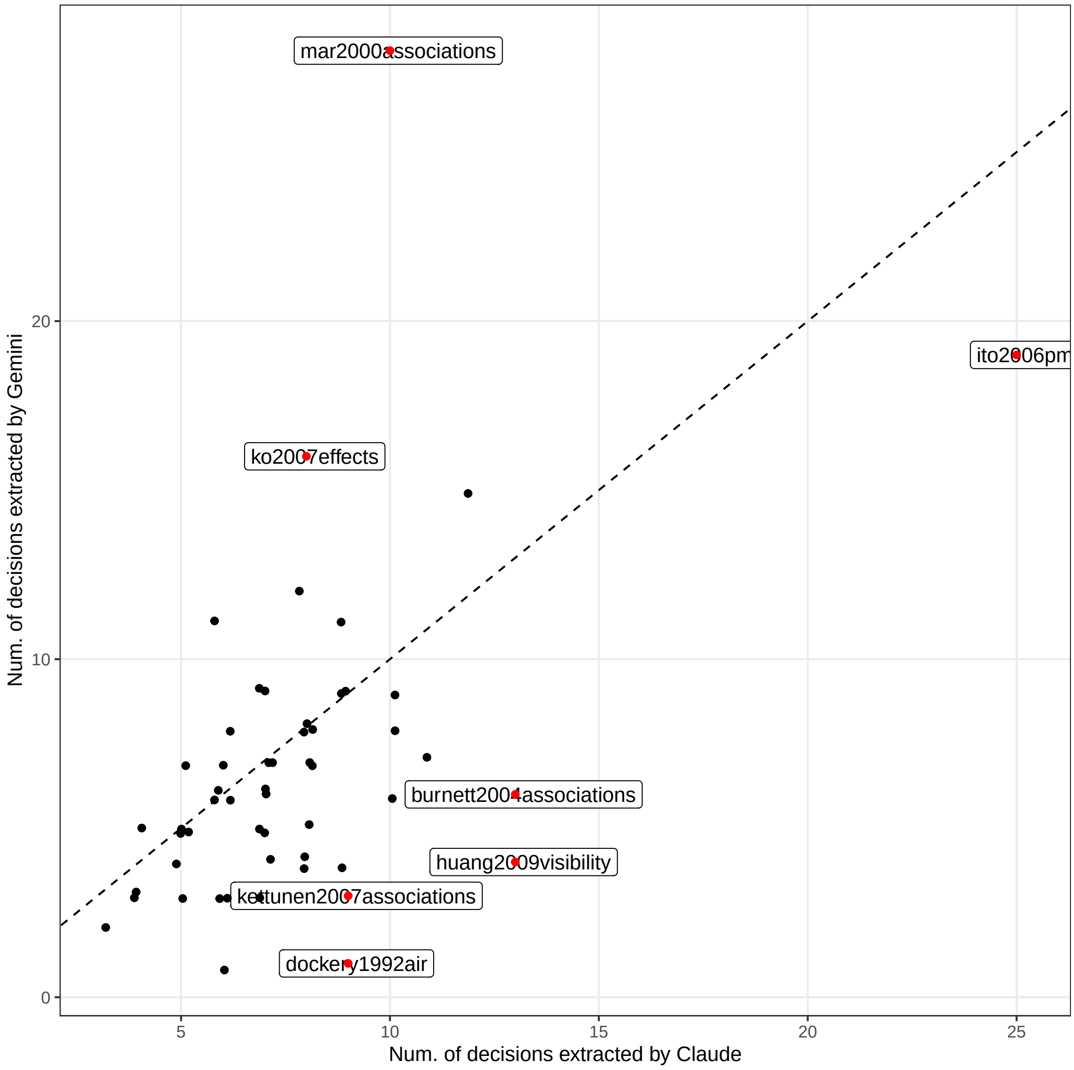

Figure 4. Comparison of decisions extracted by Claude and Gemini. Each point represents a paper, with the x- and y-axis showing the number of decisions extracted by Claude and Gemini, respectively. The dashed 1:1 line marks where both models extract the same number of decisions. More points fall below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

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. (2000) (Claude: 10 vs. Gemini: 28), Ito et al. (2006) (Claude: 25 vs. Gemini: 19), Ko et al. (2007) (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. (1992) and Burnett et al. (2004) for Gemini and Dockery et al. (1992) and Katsouyanni et al. (2001) 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 (Devlin et al. 2019) by Google, 2) RoBERTa (Liu et al., n.d.) by Facebook AI, trained on a larger dataset (160GB v.s. BERT's 15GB), 3) XLNnet (Yang et al., n.d.) by Google Brain, and two domain-trained BERT models: 4) sciBERT (Beltagy et al. 2019), trained on scientific literature, and 5) bioBERT (Lee et al. 2020), 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

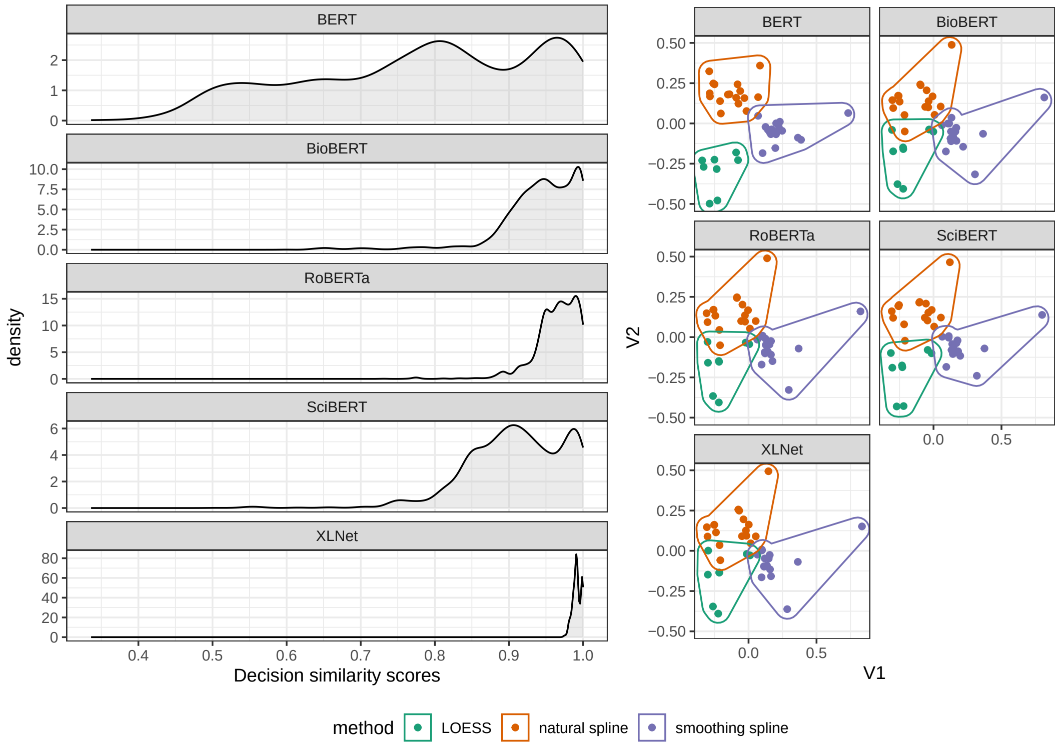

Figure 5. Distribution of decision similarity (left) and multi-dimensional scaling (MDS) of the paper similarity scores (right) computed for five different text models (BERT, BioBERT, RoBERTa, SciBERT, and XLNet). The default language model, BERT, produces the widest variation across the five models, 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).

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 (Harrod et al. 2024; Katz et al. 2024; Farzi et al. 2024; Hu et al. 2024; Sciannameo et al. 2024; Gu et al. 2025; Schilling-Wilhelmi et al. 2025; Gupta et al. 2024; Li et al. 2024; Baddour et al. 2024; Polak and Morgan 2024) 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 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. (2001) 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 (Peng et al. 2006; Touloumi et al. 2006).

7. Acknowledgement

The article has been created using Quarto (Allaire et al. 2022) in R (R Core Team 2025). The source code for reproducing the work reported in this paper can be found at: https://github.com/huizezhang-sherry/paper-decisions.

References

Alexander, Eric, Joe Kohlmann, Robin Valenza, Michael Witmore, and Michael Gleicher. 2014. "2014 IEEE Conference on Visual Analytics Science and Technology (VAST)." October, 173–82. https://doi.org/10.1109/VAST.2014.7042493.

Allaire, J. J., C. Teague, C. Scheidegger, Y. Xie, and C. Dervieux. 2022. *Quarto*. Version 1.2. https://doi.org/10.5281/zenodo.5960048.

Alspaugh, Sara, Nava Zokaei, Andrea Liu, Cindy Jin, and Marti A. Hearst. 2019.

“Futzing and Moseying: Interviews with Professional Data Analysts on Exploration Practices.” *IEEE Transactions on Visualization and Computer Graphics* 25 (1): 22–31. https://doi.org/10.1109/TVCG.2018.2865040.

Andersen, Z. J., P. Wahlin, O. Raaschou-Nielsen, M. Ketzel, T. Scheike, and S. Loft. 2008. “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): 458–66. https://doi.org/10.1136/oem.2007.033290.

Atkinson, R. W., S. Kang, H. R. Anderson, I. C. Mills, and H. A. Walton. 2014. “Epidemiological Time Series Studies of PM2.5 and Daily Mortality and Hospital Admissions: A Systematic Review and Meta-Analysis.” *Thorax* 69 (7): 660–65. https://doi.org/10.1136/thoraxjnl-2013-204492.

Baddour, Moussa, Stéphane Paquelet, Paul Rollier, Marie De Tayrac, Olivier Dameron, and Thomas Labbe. 2024. “2024 IEEE 12th International Conference on Intelligent Systems (IS).” August, 1–8. https://doi.org/10.1109/IS61756.2024.10705235.

Beltagy, Iz, Kyle Lo, and Arman Cohan. 2019. “Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).” (Hong Kong, China), 3613–18. https://doi.org/10.18653/v1/D19-1371.

Bethard, Steven, and Dan Jurafsky. 2010. “CIKM ’10: International Conference on

Information and Knowledge Management." (Toronto ON Canada), October 26, 609–18. https://doi.org/10.1145/1871437.1871517.

Blair, Graeme, Jasper Cooper, Alexander Coppock, and Macartan Humphreys. 2019. "Declaring and Diagnosing Research Designs." *American Political Science Review* 113 (3): 838–59. https://doi.org/10.1017/S0003055419000194.

Botvinik-Nezer, Rotem, Felix Holzmeister, Colin F. Camerer, et al. 2020. "Variability in the Analysis of a Single Neuroimaging Dataset by Many Teams." *Nature* 582 (7810): 84–88. https://doi.org/10.1038/s41586-020-2314-9.

Braga, Alfésio Luís Ferreira, Antonella Zanobetti, and Joel Schwartz. 2001. "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. https://journals.lww.com/joem/fulltext/2001/11000/the_lag_structure_between_particulate_air.1.aspx.

Burnett, Richard T., Sabit Cakmak, Mark E. Raizenne, et al. 1998. "The Association Between Ambient Carbon Monoxide Levels and Daily Mortality in Toronto, Canada." *Journal of the Air & Waste Management Association* 48 (8): 689–700. https://doi.org/10.1080/10473289.1998.10463718.

Burnett, Richard T., Stieb ,Dave, Brook ,Jeffrey R., et al. 2004. "Associations Between Short-Term Changes in Nitrogen Dioxide and Mortality in Canadian Cities." *Archives of Environmental Health: An International Journal* 59 (5): 228–36. https://doi.org/

10.3200/AEOH.59.5.228-236.

Castillejos, Margarita, Borja-Aburto ,Victor H., Dockery ,Douglas W., Gold ,Diane R., and Dana. and Loomis. 2000. "Airborne Coarse Particles and Mortality." *Inhalation Toxicology* 12 (sup1): 61–72. https://doi.org/10.1080/0895-8378.1987.11463182.

Chen, Chaomei. 2006. "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): 359–77. https://doi.org/10.1002/asi.20317.

Chou, J. -K., and C. -K. Yang. 2011. "PaperVis: Literature Review Made Easy." *Computer Graphics Forum* 30 (3): 721–30. https://doi.org/10.1111/j.1467-8659.2011.01921.x.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "NAACL-HLT 2019." Edited by Jill Burstein, Christy Doran, and Thamar Solorio. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423.

Dockery, Douglas W., Joel Schwartz, and John D. Spengler. 1992. "Air Pollution and Daily Mortality: Associations with Particulates and Acid Aerosols." *Environmental Research* 59 (2): 362–73. https://doi.org/10.1016/S0013-9351(05)80042-8.

Dörk, Marian, Nathalie Henry Riche, Gonzalo Ramos, and Susan Dumais. 2012. "PivotPaths: Strolling Through Faceted Information Spaces." *IEEE Transactions on Visualization and Computer Graphics* 18 (12): 2709–18. https://doi.org/10.1109/

TVCG.2012.252.

Farzi, Saeed, Soumitra Ghosh, Alberto Lavelli, and Bernardo Magnini. 2024. "Get the Best Out of 1B LLMs: Insights from Information Extraction on Clinical Documents." Edited by Dina Demner-Fushman, Sophia Ananiadou, Makoto Miwa, Kirk Roberts, and Junichi Tsujii. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.bionlp-1.21.

Gelman, Andrew, and Eric Loken. 2014. "The Statistical Crisis in Science." *American Scientist* 102 (6): 460–65. https://www.proquest.com/docview/1616141998/abstract/5E050DCE82414037PQ/1.

Götz, Martin, Abhraneel Sarma, and Ernest H. O'Boyle. 2024. "The Multiverse of Universes: A Tutorial to Plan, Execute and Interpret Multiverses Analyses Using the R Package Multiverse." *International Journal of Psychology* 59 (6): 1003–14. https://doi.org/10.1002/ijop.13229.

Gould, Elliot, Hannah S. Fraser, Timothy H. Parker, et al. 2025. "Same Data, Different Analysts: Variation in Effect Sizes Due to Analytical Decisions in Ecology and Evolutionary Biology." *BMC Biology* 23 (1): 35. https://doi.org/10.1186/s12915-024-02101-x.

Gu, Bowen, Vivian Shao, Ziqian Liao, et al. 2025. "Scalable Information Extraction from Free Text Electronic Health Records Using Large Language Models." *BMC Medical Research Methodology* 25 (1): 23. https://doi.org/10.1186/s12874-025-02470-z.

Gu, Ken, Eunice Jun, and Tim Althoff. 2023. "Understanding and Supporting Debugging Workflows in Multiverse Analysis." (New York, NY, USA), CHI '23, April 19, 119. https://doi.org/10.1145/3544548.3581099.

Gupta, Sonakshi, Akhlak Mahmood, Pranav Shetty, Aishat Adeboye, and Rampi Ramprasad. 2024. "Data Extraction from Polymer Literature Using Large Language Models." *Communications Materials* 5 (1): 269. https://doi.org/10.1038/s43246-024-00708-9.

Harrod, Karlyn K., Prabin Bhandari, and Antonios Anastasopoulos. 2024. "From Text to Maps: LLM-Driven Extraction and Geotagging of Epidemiological Data." Edited by Daryna Dementieva, Oana Ignat, Zhijing Jin, et al. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.nlp4pi-1.24.

Heimerl, Florian, Qi Han, Steffen Koch, and Thomas Ertl. 2016. "CiteRivers: Visual Analytics of Citation Patterns." *IEEE Transactions on Visualization and Computer Graphics* 22 (1): 190–99. https://doi.org/10.1109/TVCG.2015.2467621.

Hu, Yan, Qingyu Chen, Jingcheng Du, et al. 2024. "Improving Large Language Models for Clinical Named Entity Recognition via Prompt Engineering." *Journal of the American Medical Informatics Association* 31 (9): 1812–20. https://doi.org/10.1093/jamia/ocad259.

Huntington-Klein, Nick, Andreu Arenas, Emily Beam, et al. 2021. "The Influence of Hidden Researcher Decisions in Applied Microeconomics." *Economic Inquiry* 59 (3):

944–60. https://doi.org/10.1111/ecin.12992.

Isenberg, Petra, Tobias Isenberg, Michael Sedlmair, Jian Chen, and Torsten Möller. 2017. "Visualization as Seen Through Its Research Paper Keywords." *IEEE Transactions on Visualization and Computer Graphics* 23 (1): 771–80. https://doi.org/10.1109/TVCG.2016.2598827.

Ito, Kazuhiko, William F. Christensen, Delbert J. Eatough, et al. 2006. "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. https://doi.org/10.1038/sj.jea.7500464.

Jun, Eunice, Melissa Birchfield, Nicole De Moura, Jeffrey Heer, and René Just. 2022. "Hypothesis Formalization: Empirical Findings, Software Limitations, and Design Implications." *ACM Transactions on Computer-Human Interaction (TOCHI)* 29 (1): 1–28.

Jun, Eunice, Maureen Daum, Jared Roesch, et al. 2019. "Tea: A High-Level Language and Runtime System for Automating Statistical Analysis." (New York, NY, USA), UIST '19, October 17, 591603. https://doi.org/10.1145/3332165.3347940.

Jun, Eunice, Audrey Seo, Jeffrey Heer, and René Just. 2022. "Tisane: Authoring Statistical Models via Formal Reasoning from Conceptual and Data Relationships." (New York, NY, USA), CHI '22, April 29, 116. https://doi.org/10.1145/3491102.3501888.

Kale, Alex, Matthew Kay, and Jessica Hullman. 2019. “Decision-Making Under Uncertainty in Research Synthesis: Designing for the Garden of Forking Paths.” (New York, NY, USA), CHI ’19, May 2, 114. https://doi.org/10.1145/3290605.3300432.

Kale, Alex, Sarah Lee, Terrance Goan, Elizabeth Tipton, and Jessica Hullman. 2023. “MetaExplorer : Facilitating Reasoning with Epistemic Uncertainty in Meta-Analysis.” (New York, NY, USA), CHI ’23, April 19, 114. https://doi.org/10.1145/3544548.3580869.

Kan, Haidong, Stephanie J. London, Guohai Chen, et al. 2007. “Differentiating the Effects of Fine and Coarse Particles on Daily Mortality in Shanghai, China.” *Environment International* 33 (3): 376–84. https://doi.org/10.1016/j.envint.2006.12.001.

Kandel, Sean, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer. 2012. “Enterprise Data Analysis and Visualization: An Interview Study.” *IEEE Transactions on Visualization and Computer Graphics* 18 (12): 2917–26. https://doi.org/10.1109/TVCG.2012.219.

Katsouyanni, Klea, Jonathan M. Samet, H. Ross Anderson, et al. 2009. *Air Pollution and Health: A European and North American Approach (APHENA)*. Research Report. (Boston, MA), no. 142.

Katsouyanni, Klea, Joel Schwartz, Claudia Spix, et al. 1996. “Short Term Effects of Air Pollution on Health: A European Approach Using Epidemiologic Time Series Data: The APHEA Protocol.” *Journal of Epidemiology & Community Health* 50 (Suppl 1):

S12–18.

Katsouyanni, Klea, Giota Touloumi, Evangelia Samoli, et al. 2001. "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. https://journals.lww.com/epidem/fulltext/2001/09000/confounding_and_effect_modification_in_the.11.aspx.

Katz, Uri, Mosh Levy, and Yoav Goldberg. 2024. "Findings of the Association for Computational Linguistics: EMNLP 2024." (Miami, Florida, USA), 8838–55. https://doi.org/10.18653/v1/2024.findings-emnlp.516.

Kjell, Oscar, Salvatore Giorgi, and H. Andrew Schwartz. 2023. "The Text-Package: An r-Package for Analyzing and Visualizing Human Language Using Natural Language Processing and Deep Learning." *Psychological Methods*, ahead of print. https://doi.org/10.1037/met0000542.

Ko, F. W. S., W. Tam, T. W. Wong, et al. 2007. "Effects of Air Pollution on Asthma Hospitalization Rates in Different Age Groups in Hong Kong." *Clinical & Experimental Allergy* 37 (9): 1312–19. https://doi.org/10.1111/j.1365-2222.2007.02791.x.

Lafferty, John, Andrew McCallum, and Fernando Pereira. n.d. *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data.*

Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, et al. 2020. "BioBERT: A Pre-Trained

Biomedical Language Representation Model for Biomedical Text Mining." *Bioinformatics* 36 (4): 1234–40. https://doi.org/10.1093/bioinformatics/btz682.

Li, Ni, Shorouq Zahra, Mariana Brito, et al. 2024. "Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)." (Bangkok, Thailand), 93–110. https://doi.org/10.18653/v1/2024.climatenlp-1.7.

Liu, Jiali, Nadia Boukhelifa, and James R. Eagan. 2020. "Understanding the Role of Alternatives in Data Analysis Practices." *IEEE Transactions on Visualization and Computer Graphics* 26 (1): 66–76. https://doi.org/10.1109/TVCG.2019.2934593.

Liu, Yang, Tim Althoff, and Jeffrey Heer. 2020. "Paths Explored, Paths Omitted, Paths Obscured: Decision Points & Selective Reporting in End-to-End Data Analysis." (New York, NY, USA), CHI '20, April 23, 114. https://doi.org/10.1145/3313831.3376533.

Liu, Yang, Alex Kale, Tim Althoff, and Jeffrey Heer. 2021. "Boba: Authoring and Visualizing Multiverse Analyses." *IEEE Transactions on Visualization and Computer Graphics* 27 (2): 1753–63. https://doi.org/10.1109/TVCG.2020.3028985.

Liu, Yinhan, Myle Ott, Naman Goyal, et al. n.d. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. https://doi.org/10.48550/arXiv.1907.11692.

López-Villarrubia, Elena, Ferran Ballester, Carmen Iñiguez, and Nieves Peral. 2010. "Air Pollution and Mortality in the Canary Islands: A Time-Series Analysis." *Environmental Health* 9 (February): 8. https://doi.org/10.1186/1476-069X-9-8.

Mar, T F, G A Norris, J Q Koenig, and T V Larson. 2000. "Associations Between Air Pollution and Mortality in Phoenix, 1995-1997." *Environmental Health Perspectives* 108 (4): 347–53. https://doi.org/10.1289/ehp.00108347.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. "Distributed Representations of Words and Phrases and Their Compositionality." 26. https://papers.nips.cc/paper_files/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.

Moolgavkar, Suresh H. 2000. "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): 1199–206. https://doi.org/10.1080/10473289.2000.10464162.

Moolgavkar, Suresh H. 2003. "Air Pollution and Daily Mortality in Two u.s. Counties: Season-Specific Analyses and Exposure-Response Relationships." *Inhalation Toxicology* 15 (9): 877–907. https://doi.org/10.1080/08958370390215767.

Nadeau, David, and Satoshi Sekine. 2007. "A Survey of Named Entity Recognition and Classification." *Lingvisticæ Investigationes* 30 (1): 3–26. https://doi.org/10.1075/li.30.1.03nad.

Narechania, Arpit, Alireza Karduni, Ryan Wesslen, and Emily Wall. 2022. "VITALITY: Promoting Serendipitous Discovery of Academic Literature with Transformers & Visual Analytics." *IEEE Transactions on Visualization and Computer Graphics* 28

(1): 486–96. https://doi.org/10.1109/TVCG.2021.3114820.

Peng, Roger D., Francesca Dominici, and Thomas A. Louis. 2006. "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. https://doi.org/10.1111/j.1467-985X.2006.00410.x.

Polak, Maciej P., and Dane Morgan. 2024. "Extracting Accurate Materials Data from Research Papers with Conversational Language Models and Prompt Engineering." *Nature Communications* 15 (1): 1569. https://doi.org/10.1038/s41467-024-45914-8.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Samet, Jonathan M., Francesca Dominici, Frank C. Curriero, Ivan Coursac, and Scott L. Zeger. 2000. "Fine Particulate Air Pollution and Mortality in 20 u.s. Cities, 1987–1994." *New England Journal of Medicine* 343 (24): 1742–49. https://doi.org/10.1056/NEJM200012143432401.

Sarma, Abhraneel, Alex Kale, Michael Moon, et al. 2021. "Multiverse: Multiplexing Alternative Data Analyses in r Notebooks (Version 0.6.2)." *OSF Preprints*. https://github.com/MUCollective/multiverse.

Sarstedt, Marko, Susanne J. Adler, Christian M. Ringle, et al. 2024. "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. https://doi.org/10.1111/jpim.12738.

Schilling-Wilhelmi, Mara, Martiño Ríos-García, Sherjeel Shabih, et al. 2025. "From Text to Insight: Large Language Models for Chemical Data Extraction." *Chemical Society Reviews* 54 (3): 1125–50. https://doi.org/10.1039/D4CS00913D.

Schwartz, Joel. 2000. "The Distributed Lag Between Air Pollution and Daily Deaths." *Epidemiology* 11 (3): 320–26. https://www.jstor.org/stable/3703220.

Sciannameo, Veronica, Daniele Jahier Pagliari, Sara Urru, et al. 2024. "Information Extraction from Medical Case Reports Using OpenAI InstructGPT." *Computer Methods and Programs in Biomedicine* 255 (October): 108326. https://doi.org/10.1016/j.cmpb.2024.108326.

Silberzahn, R., E. L. Uhlmann, D. P. Martin, et al. 2018. "Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results." *Advances in Methods and Practices in Psychological Science* 1 (3): 337–56. https://doi.org/10.1177/2515245917747646.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11): 1359–66. https://doi.org/10.1177/0956797611417632.

Simson, Jan, Fiona Draxler, Samuel Mehr, and Christoph Kern. 2025. "Preventing Harmful Data Practices by Using Participatory Input to Navigate the Machine Learning Multiverse." (New York, NY, USA), CHI '25, April 25, 130. https://doi.org/10.1145/3706598.3713482.

Tbahriti, Imad, Christine Chichester, Frédérique Lisacek, and Patrick Ruch. 2006. "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*, Recent advances in natural language processing for biomedical applications special issue, vol. 75 (6): 488–95. https://doi.org/10.1016/j.ijmedinf.2005.06.007.

Touloumi, G., E. Samoli, M. Pipikou, A. Le Tertre, R. Atkinson, and K. Katsouyanni. 2006. "Seasonal Confounding in Air Pollution and Health Time-Series Studies: Effect on Air Pollution Effect Estimates." *Statistics in Medicine* 25 (24): 4164–78. https://doi.org/10.1002/sim.2681.

Ueda, Kayo, Nitta ,Hiroshi, Ono ,Masaji, and Ayano and Takeuchi. 2009. "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): 1212–18. https://doi.org/10.3155/1047-3289.59.10.1212.

Wicherts, Jelte M., Coosje L. S. Veldkamp, Hilde E. M. Augusteijn, Marjan Bakker, Robbie C. M. van Aert, and Marcel A. L. M. van Assen. 2016. "Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking." *Frontiers in Psychology* 7 (November). https://doi.org/10.3389/fpsyg.2016.01832.

Wickham, Hadley. 2014. "Tidy Data." *Journal of Statistical Software* 59 (September): 1–23. https://doi.org/10.18637/jss.v059.i10.

Wickham, Hadley, Mara Averick, Jennifer Bryan, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Joe Cheng, and Aaron Jacobs. 2025. *Ellmer: Chat with Large Language Models*. https://CRAN.R-project.org/package=ellmer.

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. n.d. *XLNet: Generalized Autoregressive Pretraining for Language Understanding*. https://doi.org/10.48550/arXiv.1906.08237.