

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

3   **ANONYMOUS AUTHOR(S)**

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

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

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

16   **ACM Reference Format:**

17   Anonymous Author(s). 2025. An LLM-based pipeline for understanding decision choices in data analysis from published literature.  
18   In *Proceedings of CHI Conference on Human Factors in Computing Systems (CHI'26)*. ACM, New York, NY, USA, 24 pages. <https://doi.org/XXXXXXX.XXXXXXX>

19   **1 Introduction**

20   TODO: need references

21   Decisions are made at every stage of data analysis, from initial data collection, pre-processing to modeling. One  
22   might expect well-trained researchers to make similar choices when faced with the same analytical task, yet evidence  
23   suggests otherwise. Many-analyst experiments show that independent analysts often arrive at markedly different  
24   conclusions, even when analyzing the same dataset to answer the same research question [8, 21, 67]. This variation in  
25   analytical decision-making, described by Gelman and Loken [20] as the “garden of forking paths,” can undermine the  
26   quality and credibility of reported results and hinder comparability across studies. For junior researchers who lack  
27   guidance, this variability may lead to over reliance on default statistical software settings or arbitrary choices made  
28   without clear justification.

29   A common approach to investigate uncertainty in analytical decisions is sensitivity analysis, where researchers  
30   systematically vary key decisions in their analysis to assess the robustness of their findings. Multiverse analysis extends  
31   this idea by evaluating *all* plausible combinations of decision choices to examine how results vary across the full  
32   decision space [7, 62]. However, what an analyst consider reasonable may not reflect the full range of options used in

33   

---

34   Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not  
35   made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components  
36   of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to  
37   redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

38   © 2025 ACM.

39   Manuscript submitted to ACM

40   Manuscript submitted to ACM

practice. Even when a reasonable set of alternatives is tested, the sensitivity analysis may be of limited interest to other researchers facing a similar problem, who are seeking evidence to inform comparable decision. Ideally, decision-making in applied research would 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. Classic research training typically involves reading through the literature to learn the common choices and to understand how decisions are made. This process now has the possibility to be automated at scale given recent advance in information extraction with Large Language Models’ (LLMs) [4, 19, 22, 24, 26, 28, 42, 46, 60, 64, 66].

In this work, we propose a new approach for 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. This type of studies is 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, such as the number of knots or degree of freedom for smoothing methods 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. We also use a secondary LLM to standardize reported choices of temporal lag decisions. Additionally, we conduct sensitivity analysis on reproducibility across runs and model providers. future studies for information extraction task with LLMs.

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 response, 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.

## 105 2 Related work

### 106 107 2.1 Analytic decision making in data analysis

108 Data analysis is a complex and iterative process [32–34] that involves multiple stages, including data collection, data  
109 cleaning, visualization, modeling, and communication. At each stage, analysts make decisions informed by domain  
110 practices, statistical knowledge, and the data. These decisions, such as which variables to include in a model, how  
111 to handle missing data, and how hyper-parameters are chosen, act as branching points in the analysis workflow.  
112 [TODO]The full set of possible paths through these branching points form what Gelman and Loken [20] describe as  
113 the “garden of forking paths”. While one might expect well-trained researchers to make similar choices when facing  
114 similar decisions, empirical evidence suggests otherwise. “Many analyst experiments” show that independent research  
115 groups analyzing the same dataset to address the same research questions can arrive at widely different conclusions.  
116 For example, Silberzahn et al. [67] asks 29 groups of analysts to conduct an analysis to address the same research  
117 questions *whether soccer players with dark skin tone are more likely than those with light skin tone to receive red cards*  
118 *from referees*. Researchers reported an estimated effect size from 0.89 to 2.93 in odds ratio with 21 unique combinations  
119 of covariates are used among all 29 analyses. 70% of the teams found a statistically significant positive effect while  
120 others don’t. This great discrepancy among researchers when performing data analysis task is also observed in other  
121 domains, for example, structural equation modeling [63], applied microeconomics [30], neuroimaging [8], and ecology  
122 and evolutionary biology [21].

123 Examples like the above illustrate how analytical decisions introduce uncertainty into data analysis. These uncer-  
124 tainties have been widely discussed in the literature given their impact for policy recommendation [41] and [TODO]  
125 applications in health, finance, fairness machine learning [69]. Through experiments, research has shown that analysts’  
126 decisions can lead to p-hacking and inflated effect size, when not properly used [68, 73]. Hence, guidelines and check-  
127 lists have been developed to recommend the best practice to guide statistical analysis. In medicine and biostatistics,  
128 pre-registration is a common practice to regulate analysts making decisions after seeing the data [? ]. Given the nuanced  
129 nature of data analysis, more work have examined how analysts make decisions in practice through interviews in both  
130 academia and industry. These studies include qualitative analysis of the decisions made [35, 48], interviews with data  
131 analysts about exploratory data analysis practice in industry [2, 38] and about how they consider alternatives in data  
132 analysis [47].

133 In addition to qualitative studies, software tools have developed to help researchers account for alternatives and  
134 uncertainties and make informed decisions in data analysis. Examples include Tea [32], which support general statistical  
135 analysis; Tisane [34], which guides choices in generalized linear mixed-effects models (GLMMs); and MetaExplore  
136 [36], which account for epistemic uncertainty (decision uncertainty) in meta-analysis. The DeclareDesign package  
137 [7] proposes the MIDA framework for researchers to declare, diagnose, and redesign their analyses with account for  
138 uncertainties of reporting the statistic of interest. Multiverse analysis proposes a different method to allow researchers  
139 to evaluate *all* plausible combinations of decision choices to examine how results vary in the full decision space. Work  
140 has been done on the software tools to support multiverse analysis [25, 62] and visualization of multiverse results [49],  
141 and debugging tools [23].

## 157      2.2 Automatic information extraction with LLMs

158  
 159 In natural language processing, information extraction is a task focus on extracting structured information from  
 160 unstructured text. Earlier approaches in information extraction tasks relied on rule-based systems and regular expres-  
 161 sions. More recent advances, including conditional random fields [44], word embeddings such as word2vec [53], and  
 162 transformer-based architectures like BERT [16], have led to the current use of LLM to extract information with prompts.  
 163 Using LLMs to extract unstructured text offers the advantage of automating the process at scale. Applications have  
 164 been seen in epidemiology data [26], scientific literature [42], clinical data [19, 22, 28, 66], chemistry knowledge [64],  
 165 and polymer science [24], climate extreme impact [46], phenotypes [4], and material properties [60]. An easier task in  
 166 information extraction is called Named Entity Recognition (NER) to identify short span information (1-4 tokens) like  
 167 person names and locations from unstructured text [56]. An example of this is extracting patient’s information and  
 168 vitals in clinical data. Extracting decisions from published literature is a more general task than NER, since justification  
 169 of a decision typically spans more than just a few words. Our task also requires linking information across sentences,  
 170 sometimes sections, to correctly identify the variables a decision refers to.  
 171

## 174      2.3 Visualization on scientific literature

175 With the growing volume of scientific publications and the difficulty of navigating the literature, there is an increasing  
 176 interest in developing systems to visualize and recommend scientific papers. These systems link papers based on  
 177 their similarity and relevance, typically determined by keywords [31], citation information [14], e.g. citation list and  
 178 co-citation, or combinations with other relevant paper metadata [6, 15, 18, 27], e.g. author and title. Recent approaches  
 179 incorporate text-based information using topic modeling [1], argumentation-based information retrieval [70], and  
 180 text embedding [57]. While metadata and high-level text-based information are useful for finding relevant papers,  
 181 researchers also need tools that help them *make sense* of the literature rather than simply *locating* it. In applied data  
 182 analysis, one interest is to understand how studies differ or align in their decision choices. Capturing the decision  
 183 choices and reasons that justify the choices from analyses enables the calculation of similarity among papers and can  
 184 pipe into dimension reduction methods and visualization for a global view of analysis practice in the field or recommend  
 185 similar papers based on decision similarities.  
 186

## 187      3 Methods

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

### 200      3.1 Record decisions in data analysis

201 In the study of the health effects of outdoor air pollution, one area of interest is the association between short-term,  
 202 day-to-day changes in particulate matter air pollution and daily mortality counts. This question has been studied  
 203 extensively by researchers across the globe and it serves to provide scientific evidence in the US to guide public policy  
 204 on setting the National Ambient Air Quality Standards (NAAQS) for air pollutants. While individual modeling choices  
 205 vary, these studies often share a common structure: they adjust for meteorological covariates, such as temperature and  
 206 Manuscript submitted to ACM

209 humidity, include lagged variables to account for temporal correlations, and estimate the effect size by city or region  
 210 before pooling the results with random effect. This naturally forms a “many-analyst” experiment setting to analyze  
 211 decisions in air pollution mortality modelling.  
 212

213 Consider the following excerpt from Ostro et al. [58] modeling the association between daily counts of mortality and  
 214 ambient particulate matter (PM10):  
 215

216 Based on previous findings reported in the literature (e.g., Samet et al. 2000), the basic model included a  
 217 smoothing spline for time with 7 degrees of freedom (df) per year of data. This number of degrees of  
 218 freedom controls well for seasonal patterns in mortality and reduces and often eliminates autocorrelation.  
 219

This sentence encode the following components of a decision:

- 220 • **variable**: time
- 221 • **method**: smoothing spline
- 222 • **parameter**: degree of freedom (df)
- 223 • **reason**: Based on previous findings reported in the literature (e.g., Samet et al. 2000); This number of degrees of  
 224 freedom controls well for seasonal patterns in mortality and reduces and often eliminates autocorrelation.
- 225 • **decision**: 7 degrees of freedom (df) per year of data

226 This decision can be recorded in a tabular format following the tidy data principle [74], which states that each  
 227 variable forms a column and each observation forms in a row. For our purpose, each row represents a decision made in  
 228 a paper and an analysis often include multiple decisions. We extract the original text in the paper, without paraphrase  
 229 or summarization. The decision above is a parameter choice of a statistical method applied to the variable *time*. A data  
 230 analysis may also include other types of decisions, such as temporal or spatial ones, for example, the choice of lagged  
 231 exposure for certain variables or whether the model is estimated collectively or separated for individual locations. These  
 232 decisions don’t have a specific method or parameter fields, but should still include variable, type (spatial or temporal),  
 233 reason, and decision fields.  
 234

235 Given the writing style of authors, multiple decisions may be combined in one sentence and certain fields may be  
 236 omitted. Consider a different excerpt from Ostro et al. [58]:  
 237

238 Other covariates, such as day of the week and smoothing splines of 1-day lags of average temperature  
 239 and humidity (each with 3 df), were also included in the model because they may be associated with  
 240 daily mortality and are likely to vary over time in concert with air pollution levels.  
 241

242 This sentence contains four decisions: two for temperature (the temporal lag and the smoothing spline parameter)  
 243 and two for humidity, and should be structured as separate entries:  
 244

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 scenario 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 document relies on Optical Character Recognition (OCR) to convert scanned images into machine-readable text – a capability currently offered by Anthropic Claude and Google Gemini. In the prompt, we assign the LLM a role as an applied statistician and instruct it 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 of reason and decision fields as discussed in Section 3.1. Prompt engineering techniques [13, 77] 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 [76] 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 ensure different expressions of the same variable are standardized into the same expression. For example, the expression *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 need to be confirmed by pressing the “Apply changes” button to update to the full dataset. The corresponding `tidyverse` [75] 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)

Edit decision table output									
Initial view									
paper	id	model	variable	method	parameter	type	reason	decision	reference
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)	
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
Upon confirmation, the changes will be applied to the full dataset									
paper	id	model	variable	method	parameter	type	reason	decision	reference
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.

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 similarities 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.

### 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 these literature. The outputs from LLM need to be validated and standardized before further analysis, due to 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.

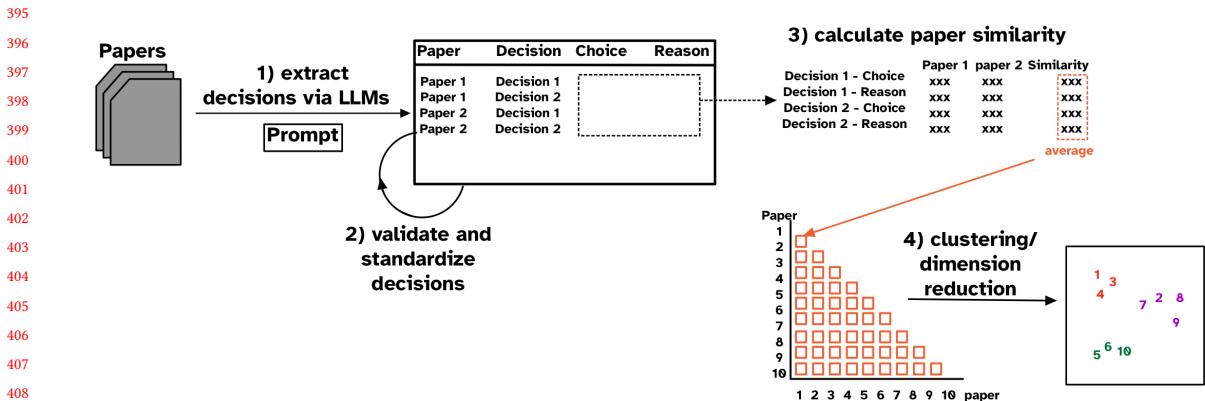


Fig. 2. The workflow for extracting decisions from published literature using Large Language Models (LLMs) and analyzing the extracted decisions. The workflow consists of four main steps: (1) Extract decisions automatically from literature with LLMs, (2) Validate and standardize LLM outputs, (3) Calculate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

417 **4 Results**

418 From the 56 studies examining the effect of particulate matters ( $PM_{10}$  and  $PM_{2.5}$ ) on mortality and hospital admission,  
 419 we focus on the baseline model reported in each paper, excluding secondary models (e.g. lag-distributed models) and  
 420 sensitivity analysis. We also exclude decisions on other pollutants, such as nitrogen dioxide ( $NO_2$ ). This yields 242  
 421 decisions extracted using Gemini, averaging approximately 4 decisions per paper.  
 422

424 **4.1 Validation and standardization of LLM outputs**

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

430 Reason	431 Count
432 Remove decisions out of scope: other pollutants and sensitivity analysis	50
433 Edit made to recode smoothing parameter unit to per year	45
434 Duplicates	9
435 Fix incorrect capture	9
436 Edit made due to decisions are too general, e.g. minimum of 1 df per year was required	6
437 Remove decisions related to definition of variables, e.g. season	5
438 Total	124

441  
 442  
 443 Table 2 summarizes the number of edits made during the review process using the Shiny application. These edits  
 444 fall into two main categories: 1) correcting LLM outputs and 2) standardizing extracted decision. The first category  
 445 includes fixing incorrect captures, removing non-decision (e.g. definition of variables), removing duplication, excluding  
 446 irrelevant decisions (e.g. sensitivity analyses), and excluding decisions whose stated reasons reflect general guidelines  
 447 rather than actual choices (e.g. “minimum of 1 degree of freedom per year is required”).  
 448

449 Standardization addresses variation in how authors express variable names and decisions. For example, variable  
 450 names such as “mean temperature” and “average temperature” refer to the same variable and should be aligned for  
 451 comparison for later decision similarity calculation. Variable names are manually standardized into four main categories:  
 452

- 453 • **temperature**: “mean temperature”, “average temperature”, “temperature”, “air temperature”, “ambient tempera-  
 454 ture”
- 455 • **humidity**: “dewpoint temperature” and its hyphenated variants, relative humidity”, “humidity”
- 456 • **PM**: “pollutant”, “pollution”, “particulate matter”, “particulate”, “PM10”, “PM2.5”
- 457 • **time**: “date”, “time”, “trends”, “trend”

458 Notice that “dewpoint temperature” is standardized under humidity because it serves as a proxy for temperature in  
 459 achieving a 100% relative humidity.  
 460

461 Decisions themselves also require standardization. For example, the smoothing parameter (number of knots and  
 462 degree of freedom) may be expressed *per year* or *in total*, and temporal lag decision may be expressed in different  
 463 formats (e.g. “6-day average”, “mean of lags 0+1”, “lagged exposure up to 6 days”). Smoothing parameter units are  
 464 manually recoded to a *per year* basis for consistency, as reflected in Table 2. Temporal decision show a wider variety,  
 465 generally falling into two categories:  
 466

469 Table 3. Missingness of decision and reason fields in the Gemini-extracted decisions. Most decisions report the choice (35.5 + 57.1 =  
 470 92%), but 57.1% lacks a stated reason.

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

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

This variability makes manual standardization impractical, hence we apply a secondary LLM process (claude-3-7-sonnet-latest) using the ellmer package to convert temporal decisions into a consistent format: `multi-day: lag [start]-[end]` and `single-day: lag [start], . . . , lag [end]`. For instance, “6-day average” is converted to “multi-day: lag 0-5” and “lagged exposure up to 6 days” is converted to “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 reason for the extracted decisions. While 2% of decisions are complete for both decision and reasons, 55% of decisions lack a stated rationale for the choice. This reflects a common reporting practice in the field, where authors often present the decision itself without providing a justification, e.g. “We decide to use  $x$  degree of freedom for variable  $y_1$  and  $y_2$ ”. This also includes cases where authors provide general guidelines for selecting the parameter, but the rationale is too broad to justify the specific choice made (hence validated as NA in Section 4.1).

500  
 501 Table 4. Count of variable-type decisions in the Gemini-extracted decisions. The most commonly reported decision are the parameter  
 502 choices and temporal lags 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

517 Table 4 lists the eight most frequently reported decision: parameter and temporal choice for time, PM, temperature,  
 518 and humidity. While a wider list of variables have been used in the analysis, these four variables are most commonly  
 519  
 520 Manuscript submitted to ACM

521 included in baseline models. Parameter choices for time, temperature, and humidity are typically made on the use of  
 522 smoothing parameter for the smoothing method (natural spline and smoothing spline), whereas temporal choices are  
 523 commonly reported for PM, temperature, and humidity for the number of lag to consider in the model.  
 524

525  
 526 Table 5. Options captured for parameter choices for time, humidity, and temperature variables in the Gemini-extracted decisions.  
 527 The choices for natural spline knots are generally less varied than the degree of freedom choices for smoothing spline. Choices for  
 528 temperature and humidity tend to be close, given they are both weather related variables, while the choices for time are more varied  
 529 inherently.

530

531 Method	532 Variable	533 Decision
533 natural spline	humidity	3, 4
534 natural spline	temperature	3, 4, 6
535 natural spline	time	1, 1.5, 3, 4, 6, 7, 8, 12, 15, 30
537 smoothing spline	humidity	2, 3, 4, 6, 8, 50% of the data
538 smoothing spline	temperature	2, 3, 4, 6, 8, 50% of the data
539 smoothing spline	time	1, 3, 4, 5, 6, 7, 7.7, 8, 9, 10, 12, 30, 100, 5% of the data

541  
 542 Table 5 presents the parameter-related decisions extracted for spline methods (natural and smoothing spline) applied  
 543 to variable time, humidity and temperature. These decisions concern the number of knots or degree of freedom, with all  
 544 values standardized to a *per year* scale for consistency. The selection of knot for natural spline has less variation than  
 545 the degree of freedom choices for smoothing spline. Choices for temperature and humidity are generally similar, given  
 546 they are both weather related variables, whereas choices for time are more varied. This tabulation provides a reference  
 547 set for common parameter choices for future studies and help to identify anomalies and special treatment in practice.  
 548 For example, the choice of 7.7 degree of freedom reported in Castillejos et al. [12] may prompt analysts to seek further  
 549 justification. By cross comparing with other reporting, some decisions appear ambiguous. For example, in Moolgavkar  
 550 [54] and Moolgavkar [55], the reported value of 30 and 100 degrees of freedom for time may be understandable for  
 551 experienced domain researcher, it could be unclear for junior analysts as to whether they apply to the full 9 year period  
 552 or on a per-year basis. We also observe a different report style from Schwartz [65], where smoothing spline parameters  
 553 are expressed as a proportion of the data (“5% of the data” and “5% of the data”) rather than fixed numerical value.  
 557

558  
 559 Table 6. Options captured for temporal lag choices for PM, temperature, and humidity variables in the Gemini-extracted decisions.  
 560 Both single-day lags and multi-day average lags are commonly used, generally considering up to five days prior (lag 5).  
 561

562 Lag type	563 Variable	564 Decision
564 multi-day average	PM	lag 0-1, 0-2, 0-3, 0-4, 0-5, 0-6
565 multi-day average	humidity	lag 0-1, 0-2, 0-3, 0-5, 1-5, 2-4
566 multi-day average	temperature	lag 0-1, 0-2, 0-3, 0-5, 2-4
568 single-day lag	PM	lag 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13
569 single-day lag	humidity	lag 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13
570 single-day lag	temperature	lag 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13

573 Similarly, Table 6 summarizes the temporal lag choices for PM, temperature, and humidity. For single-day lags, the  
 574 lags are considered up to 13 days (approximately two weeks). For multi-day averages, 3-day and 5-day averages are  
 575 most common, although other choices such as 2-4 day average are also observed as in López-Villarrubia et al. [51]:  
 576

577 In particular, lags 0 to 1 and lags 2 to 4 averages of temperature, relative humidity, and barometric  
 578 pressure were considered as meteorological variables.  
 579

### 580 4.3 Paper similarity and clustering

581 Given the number of decisions reported in Table 4, we focus on the six most common variable-type decisions for  
 582 calculating paper similarity: parameter choices for time, temperature, and humidity, and temporal lag choices for PM,  
 583 temperature, and humidity. We also restrict our analysis to papers that report at least three of these six decisions, resulting  
 584 in 48 papers for the similarity analysis. This ensures that the paper similarity metric is based on a sufficient number of  
 585 comparable decisions. We use the default text embedding model (BERT) in the text package and cosine similarity to  
 586 compute the similarity score. Sensitivity analysis on different text embedding model is checked in Section 4.4.3. Paper  
 587 similarity is then calculated as the average of decision similarity for each paper pair. The resulting distance matrix is  
 588 then used for multi-dimensional scaling (MDS) in Figure 3. The two MDS dimension reveals three clusters correspond  
 589 to the three smoothing methods used in these analyses: LOESS, natural spline, and smoothing spline. This grouping  
 590 aligns with the modeling strategies seen in large-scale analysis, such as the U.S. NMMAPS study [61] and the European  
 591 APHEA [39] and APHEA2 [40] project.  
 592

593 To reconcile these differences, the APHENA project [41] was launched with the aim to “assess the consistency across  
 594 Europe and North America when estimated using a common analytic protocol and to explore possible explanations for  
 595 any remaining variation”. While multi-dimensional scaling in Figure 3 shows the match of three clusters with three  
 596 smoothing methods, this is not inconsistent with the APHENA project [41] that the amount of smoothing to have a  
 597 more important role than the method of smoothing for estimating the effect of PM on public health variables. The  
 598 similarity metric we proposed focuses on the variation of choices across analyses, without directly assessing how those  
 599 choices influence results. By pooling decision choices from multiple studies with LLMs, it becomes much easier to  
 600 reveal common practices and difference in research practices, highlighting decisions that require further sensitivity  
 601 analyses to assess their impact. The different smoothing methods revealed in Figure 3 are consistent with the analysis  
 602 by Peng et al. [59] and Touloumi et al. [71] that compares different smoothing methods and rationale for selecting  
 603 smoothing parameters.  
 604

### 605 4.4 Sensitivity analysis

606 A series of sensitivity analysis has been conducted to explore the reproducibility for using LLMs for text extraction  
 607 tasks (Section 4.4.1), discrepancies in decision extraction between different LLM models: Gemini (gemini-2.0-flash)  
 608 and Claude (claude-3-7-sonnet-latest) (Section 4.4.2), and the sensitivity of text model for computing the semantic  
 609 decision similarity (Section 4.4.3).  
 610

611 **4.4.1 LLM reproducibility.** We assess the reproducibility of Gemini’s text extraction (gemini-2.0-flash) by repeating  
 612 the task five times for each of the 62 papers and perform pairwise comparison between runs. This generates  $5 \times 4 / 2 \times 62 =$   
 613 620 possible comparisons for both “reason” and “decisions” fields. Comparisons where the runs produced a different  
 614 number of decisions were excluded, as this would require manual alignment. After filtering, 449 out of 620 (72%)  
 615

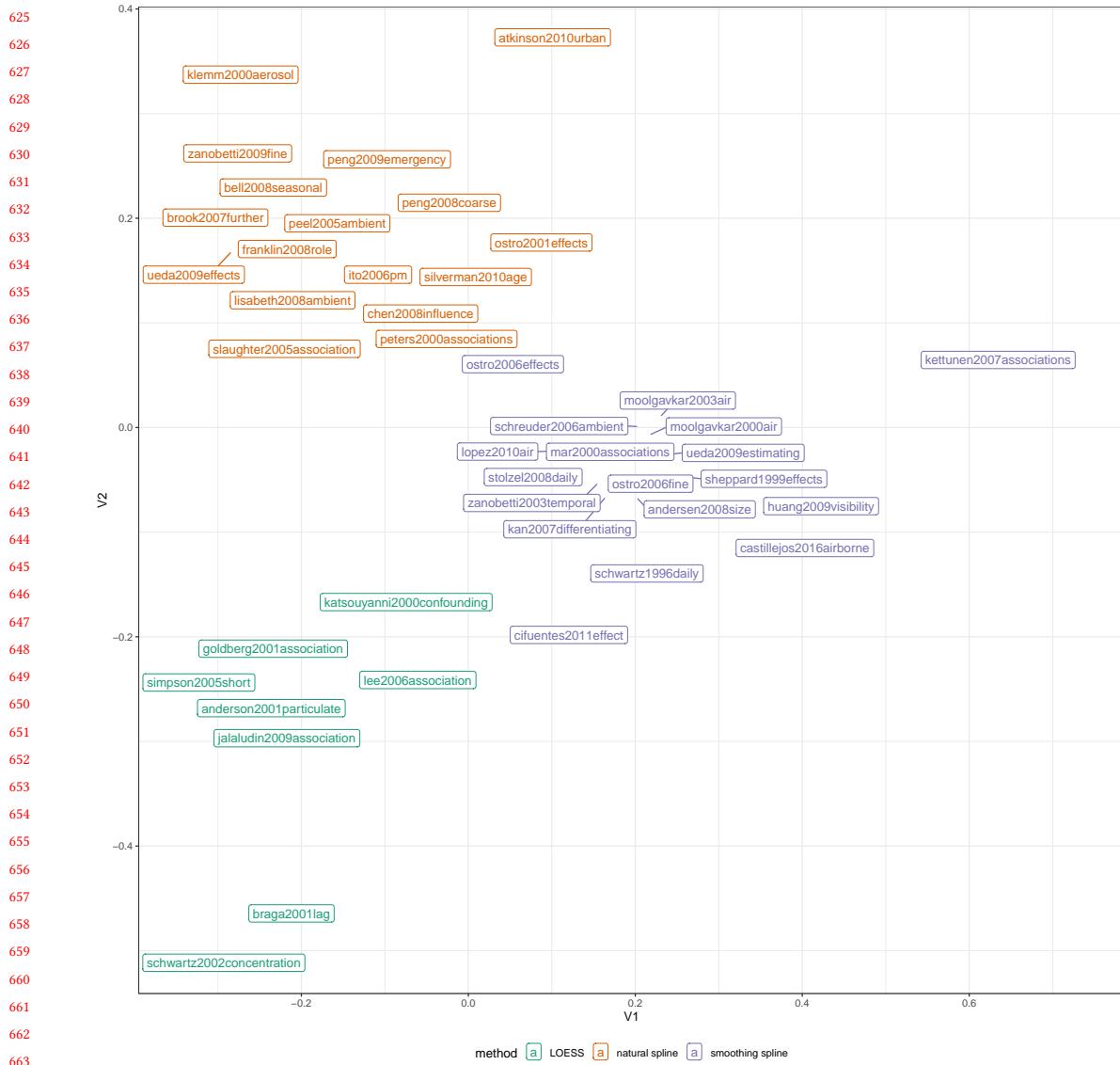


Fig. 3. The multi-dimensional scaling (MDS) based on paper similarity distance for `length(good_pp)` air pollution mortality modeling papers, colored by the smoothing method used. The MDS reveals the three distinct groups of papers, corresponds to LOESS, natural spline, and smoothing spline. These groups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

remained. Table 7 prints the decisions in Andersen et al. [3] across two runs and all the four decisions are identical with no difference.

677 Table 7. Example comparing Gemini’s text extraction for Andersen et al. [3] across two runs. The extracted decisions are identical in  
 678 both runs.  
 679

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

689 Table 8. Number of differences in the reason and decision fields across Gemini runs for papers with consistent number of decisions  
 690 across runs.  
 691

692	693 Num. of difference	694 Count	695 Proportion (%)
695	0	358	79.73
696	1	12	2.67
697	2	8	1.78
698	3	0	0.00
699	4	24	5.35
700	5	12	2.67
701	6	3	0.67
702	7	0	0.00
703	8	10	2.23
704	9	6	1.34
705	10	10	2.23
706	11	6	1.34
707	Total	449	100.00

712 Table 8 summarizes the number of differences observed in each pairwise comparison. Among all comparisons, 80%  
 713 produce the identical text in reason and decision. The discrepancies come from the following two reasons: 1) Gemini  
 714 extracted different length for the same decision, e.g. in Kan et al. [37], some runs may extract “singleday lag models  
 715 underestimate the cumulative effect of pollutants on mortality 2day moving average **of current and previous day**  
 716 **concentrations** (lag=01)”, while others extract “singleday lag models underestimate the cumulative effect of pollutants  
 717 on mortality 2day moving average (lag=01)”. Similarly, for decisions, some runs yield “10 df for total mortality”, while  
 718 other runs yield “10 df”. 2) Gemini fails to extract reasons in some runs but not others, e.g. in Burnett et al. [10], the  
 719 first run generates NAs in the reasons, but the remaining four runs are identical. In Ueda et al. [72] and Castillejos  
 720 et al. [12], runs 1 and 5 fail to extract the reasons and produce the same incomplete version, whereas runs 2, 3, and 4  
 721 produce accurate versions with reasons populated.  
 722

723 4.4.2 *LLM models*. Reading text from PDF document requires Optical Character Recognition (OCR) to convert images  
 724 into machine-readable text, which currently is only supported by Anthropic Claude (claude-3-7-sonnet-latest) and  
 725 Manuscript submitted to ACM

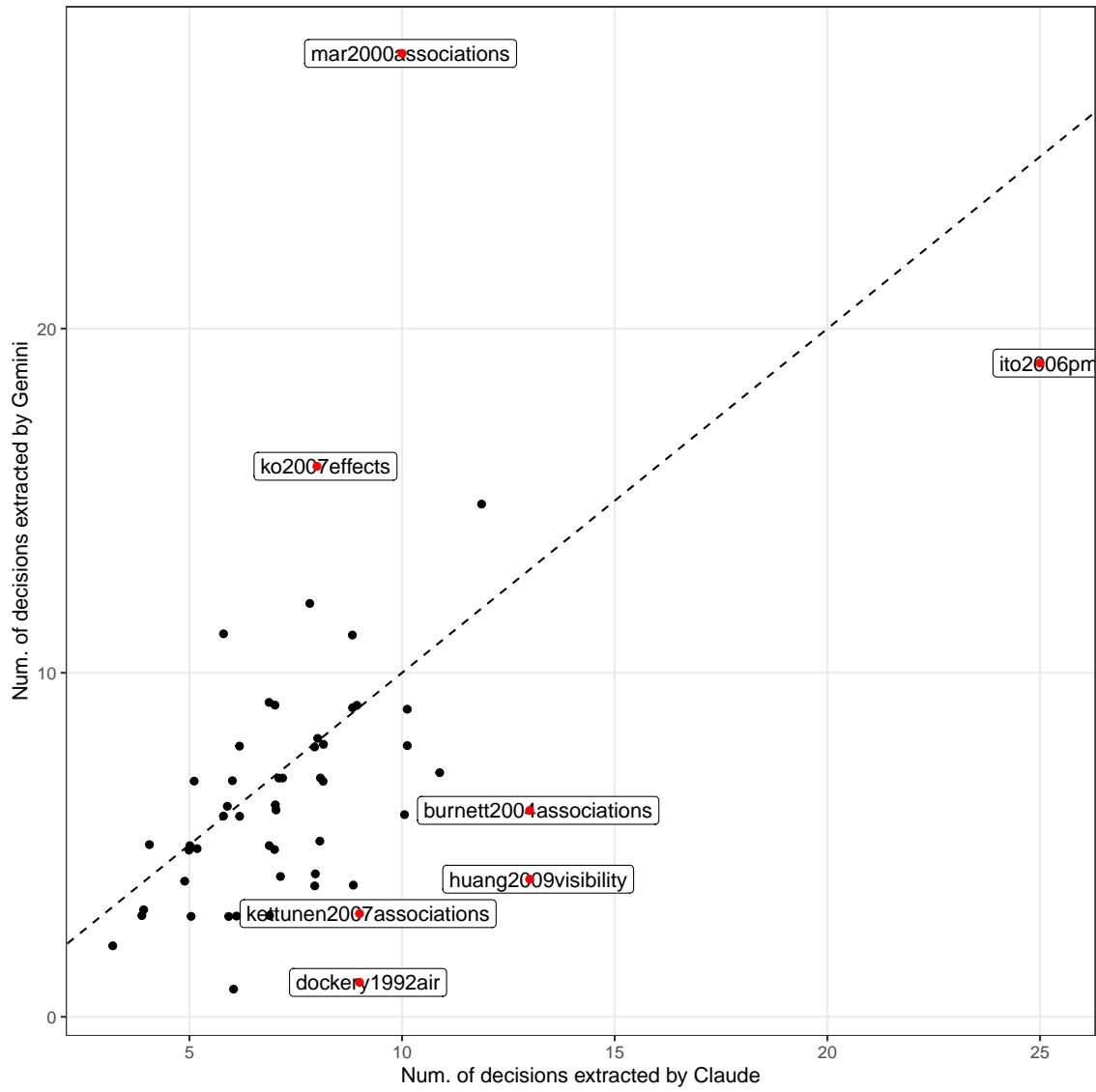


Fig. 4. Comparison of decisions extracted by Claude and Gemini. Each point represents a paper, with the x- and y-axes showing the number of decisions extracted by Claude and Gemini, respectively. The dashed 1:1 line marks where both models extract the same number of decisions. More points fall below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

Google Gemini (gemini-2.0-flash). We compare the number of decisions extracted by Claude and Gemini across all 62 papers in Figure 4. Each point represents a paper, with the x- and y-axes 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. While both models extract decisions irrelevant to our analysis, such as sensitivity analyses and secondary analyses,

Claude's extractions tend to include more of these irrelevant decisions, examples of these include 1) the definition of "cold day" and "hot day" indicators in Dockery et al. [17] ("defined at the 5th/ 95th percentile"), 2) decisions relate to other pollutants: NO<sub>2</sub>, O<sub>3</sub>, and SO<sub>2</sub> using a "24 hr average on variable" in Huang et al. [29], and 3) the definition of black smoke and in Katsouyanni et al. [40] for secondary analysis ("restrict to days with BS concentrations below 150  $\mu\text{g}/\text{m}^2$ "). While Gemini also capture these irrelevant decisions, such as "0-4 lag days" for air pollution exposure variables (CO, EC, K<sub>S</sub>, NO<sub>2</sub>, O<sub>3</sub>, OC, Pb, S, SO<sub>2</sub>, TC, Zn) in Mar et al. [52]. However, these cases are less frequent than Claude's extraction and has been validated and standardized in Section 4.1.

For both Claude and Gemini, we find they 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.

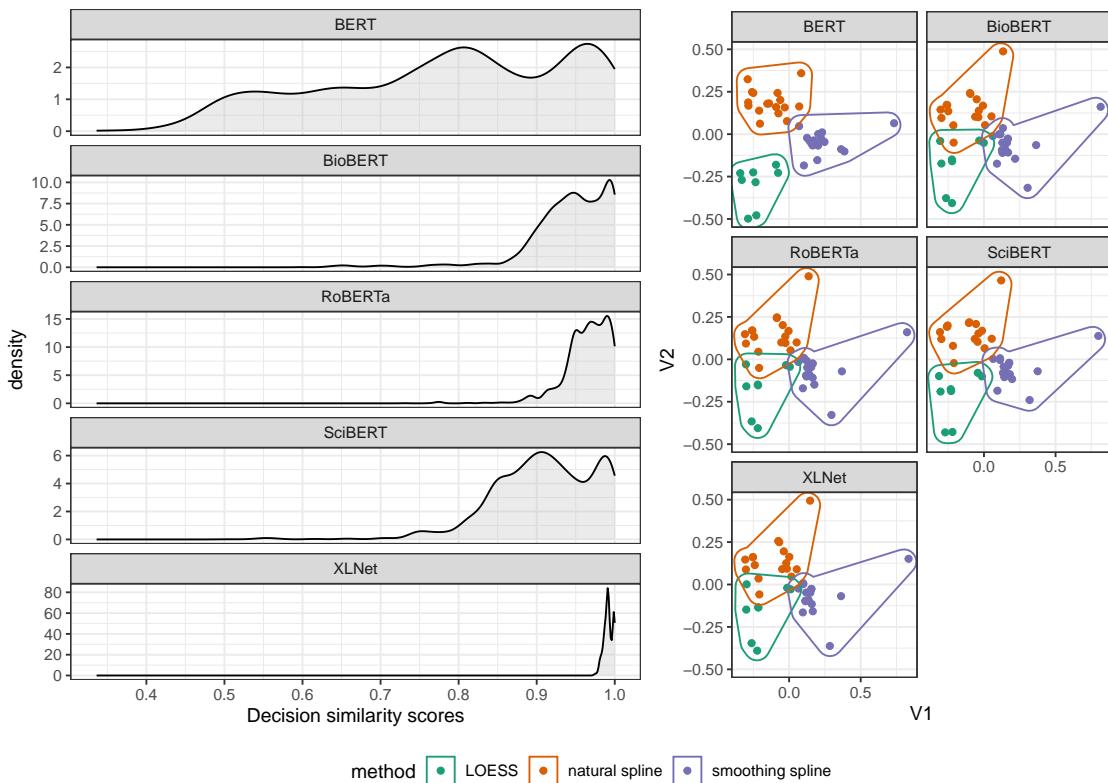


Fig. 5. Distribution of decision similarity (left) and multi-dimensional scaling (MDS) of the paper similarity scores (right) computed for five different text models (BERT, BioBERT, RoBERTa, SciBERT, and XLNet). The default language model, BERT, produces the widest variation across the five models, while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

833     4.4.3 *Text model.* We have conducted sensitivity analysis on the text model for obtaining the decision similarity score  
834     from the Gemini outputs. The tested language models tested include 1) BERT by Google [16], 2) RoBERTa by Facebook  
835     AI [50], trained on a larger dataset (160GB v.s. BERT’s 15GB), 3) XLNet by Google Brain [78], and two domain-trained  
836     BERT models: 4) sciBERT [5], trained on scientific literature, and 5) bioBERT [45], trained on PubMed and PMC data.  
837

838     Figure 5 shows the distribution of the decision similarity and the corresponding multi-dimensional scaling visualization,  
839     where distance are calcualted from the paper similarity for each text model. At decision level, the BERT model  
840     produces the widest variation across all five models, while the similarity scores from XLNet are all close to 1. While the  
841     raw scores are not directly comparable across models due to the difference in the underlying transformer architecture,  
842     the multi-dimensional scaling (MDS) based on paper similarity scores shows a similar clustering pattern corresponding  
843     to the three main smoothing methods (LOESS, natural spline, and smoothing spline).  
844

## 845     5 Discussion

### 846     5.1 Large-language models for information extraction

847     Numerous studies have demonstrated the capability of LLMs in information extraction across domains [4, 19, 22, 24, 26,  
848     28, 42, 46, 60, 64, 66]. Our work applies the LLMs to extract analytic decisions in scientific literature, providing further  
849     evidence of their effectiveness for information extraction task. Unlike named entity recognition (NER) in clinical data,  
850     our task requires capturing more complex analytical decisions and their justifications, which typically span more than  
851     just a few tokens. This also requires linking information across sentences and sometimes sections to correctly identify  
852     the variables (e.g., linking “weather” to “temperature” and “humidity”).  
853

854     While the extraction of decisions from literature could be largely automated with LLMs, manual validations remains  
855     essential to ensure the quality of the extracted decisions for downstream analysis. Most existing applications evaluate  
856     LLMs by comparing their outputs to human-annotated datasets, reporting metrics such as precision, recall, and F1 score.  
857     Because this approach depends on labeled data, which is not yet fully automated, it is not yet clear how these outputs  
858     should be validated for downstream analysis in practice. In our work, we automate some of the manual validation  
859     with a secondary LLM (Claude) to standardize the temporal lag choices into two categories: multi-day averages and  
860     single-day lags.  
861

862     With a default temperature of one and the prompt to instruct the model to extract the original text rather than  
863     paraphrase, we find that hallucination is not a major issue with Claude and Gemini in this application. Because LLM  
864     outputs are inherently probabilistic, we also conduct sensitivity analyses on reproducibility across runs and model  
865     providers. The output is generally stable: repeated runs with the Gemini produces consistent results, and different  
866     models extracted a similar number of decisions.  
867

868     While we optimize the prompt for decision extraction in this work, an alternative approach is to fine-tune a local  
869     model to enhance LLM performance. A catered local model could be useful for extraction decisions for a comprehensive  
870     literature reviews on a larger scale, but it would require greater model training efforts with labeled data.  
871

### 872     5.2 Extracting other types of decisions

873     In this work, we focus on modeling decisions for the baseline model in the air pollution epidemiology literature. Analyses  
874     in this fields often fit multiple models for different health outcomes and secondary models, such as distributed lag  
875     models and multi-pollutant models, are also commonly used to estimate relative risks and multi-pollutants interactions.  
876     These increase the complexity of decision extraction with LLMs because authors often only describe the differences  
877

885 from the baseline specification, implicitly assuming other decisions remain unchanged. Hence, LLMs will need to link  
 886 the decisions across different models and reconstruct the complete set of decisions for each model.  
 887

888 Beyond modeling choices, decisions in data pre-processing are also interesting to compare. For example, Braga et al.  
 889 [9] aggregated air pollution measures from multiple PM10 monitors within the same location into a single value. Choices  
 890 about data source, aggregation, imputation, among others, could also have impact on uncertainty of the estimation.  
 891 However, these decisions are often less well-documented in the literature than the modeling decisions, hence cannot be  
 892 extracted by LLMs. Proper documentation and reporting of these decisions in future research are needed before our  
 893 workflow could be applied to pre-processing decisions.  
 894

895 With growing advocacy for reproducibility, papers nowadays are expected to share code and data, if applicable.  
 896 Code availability provides a useful supplementary source for identifying decisions and cross-checking them against  
 897 manuscript description. However, while script may reveal what choices were made, the rationale behind these choices  
 898 is often not documented under the current practice.  
 899

### 900 901 902 5.3 Generalizability of the workflow

903 In principle, our workflow is scalable and generalizable to a random set of applied papers. However, insights about the  
 904 data analysis practices are more likely to emerge when papers share certain similarities. For example literature on the  
 905 same topic but authored by different researchers enables comparisons of practices within a field; literature that use the  
 906 same methodology across disciplines allow comparisons across fields; and literature that considers the same variables  
 907 can show how those variables are used in different domains.  
 908

909 The LLM prompt for extracting decisions will need to be customized for each application of the workflow. The  
 910 general LLM prompt structure and the data schema for recording decisions can be retained, while examples within  
 911 the prompt may be adapted to suit the specific application. The shiny application for interactively validating and  
 912 standardizing decisions can be reused across applications. Calculating paper similarity requires comparing decisions on  
 913 the same variable and type across paper pairs. For papers with limited similarities, the number of comparable decisions  
 914 may be limited. Diagnostic functions are available to display decisions side by side or provide summary statistics on the  
 915 number of comparable decisions. Uncertainty visualization could be used to highlight the confidence in the similarity  
 916 metric based on the number of comparable decisions.  
 917

918 As a new method for collecting analytic decision data from literature, our workflow can be connected to meta-analysis  
 919 to assess how different decisions influence results. More broadly, it can also be integrated into literature search and  
 920 recommender systems to suggest similar papers based on the analytic decisions they employ.  
 921

## 922 923 924 6 Conclusion

925 In this paper, we aim to study how analysts make decisions in their data analysis practice. While classic interviews  
 926 are often conducted in small scale with toy examples, we developed a pipeline for automatically extracting decisions  
 927 using LLMs (Claude and Gemini) from scientific literature. We also introduced a method for calculating paper similarity  
 928 through comparing the similarities among decisions and the similarity metric can be used as a distance to cluster  
 929 papers by their decision choices and visualization with dimension reduction algorithms, such as multidimensional  
 930 scaling. We applied this pipeline to a set of air pollution modeling literature that associates daily particulate matter  
 931 and daily mortality and hospital admission. From the extracted modeling decisions, we identify the most common  
 932 decision choices in this type of analysis and the paper similarity score calculation revealed the three clusters of paper  
 933

corresponding to different modeling strategies. These findings are all consistent with the general understanding of the field, as documented in the APHENA project [41] and other methodological comparison studies [59, 71].

While sensitivity analyses are commonly used to assess the robustness of findings to different analytical choices, the set of choices tested is often limited and selected subjectively by the authors. Our approach offers a new perspective by pooling decisions made in analyses across studies in the fields. This allows for a holistic account on the alternatives in the field and identification of both consensus and divergence within the field, providing insights for future research and methodological development.

## 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. Futzling 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. Ketzler, 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.bmjjournals.org/content/65/7/458>. Publisher: BMJ Publishing Group Ltd Section: Original article PMID: 17989204.
- [4] Moussa Baddour, Stéphane Paquette, 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](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 Barilaro, 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. Bottenthorn, 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 Iacobolla, 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. Nelson, Gustav Nilsson, 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-Porras, Emily A. Yearling, Sangsuk Yoon, Rui Yuan, Kenneth S. L. Yuen, Lei Zhang, Xu Zhang, Joshua E. Zosky, Thomas E.

- 989 Nichols, Russell A. Poldrack, and Tom Schonberg. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810):  
 990 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.  
 991 [9] Alfésio Luís Ferreira Braga, Antonella Zanobetti, and Joel Schwartz. The lag structure between particulate air pollution and respiratory and  
 992 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](https://journals.lww.com/joem/fulltext/2001/11000/the_lag_structure_between_particulate_air.1.aspx).  
 993 [10] Richard T. Burnett, Sabit Cakmak, Mark E. Raizenne, David Stieb, Renaud Vincent, Daniel Krewski, Jeffrey R. Brook, Owen Philips, and Haluk  
 994 Ozkaynak. The association between ambient carbon monoxide levels and daily mortality in toronto, canada. *Journal of the Air & Waste Management  
 995 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>.  
 996 [11] Richard T. Burnett, Stieb ,Dave , Brook ,Jeffrey R. , Cakmak ,Sabit , Dales ,Robert , Raizenne ,Mark , Vincent ,Renaud , , and Tom Dann. Asso-  
 997 ciations between short-term changes in nitrogen dioxide and mortality in canadian cities. *Archives of Environmental Health: An International  
 998 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:  
 999 <https://doi.org/10.3200/AEOH.59.5.228-236> PMID: 16201668.  
 1000 [12] Margarita Castillejos, Borja-Aburto ,Victor H. , Dockery ,Douglas W. , Gold ,Diane R. , , and Dana Loomis. Airborne coarse particles and mortality.  
 1001 *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>.  
 1002 Publisher: Taylor & Francis \_eprint: <https://doi.org/10.1080/0895-8378.1987.11463182>.  
 1003 [13] Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. Unleashing the potential of prompt engineering for large language models.  
 1004 *Patterns*, 6(6):101260, 06 2025. doi: 10.1016/j.patter.2025.101260. URL <https://www.sciencedirect.com/science/article/pii/S2666389925001084>.  
 1005 [14] Chaomei Chen. Citespac ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society  
 1006 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>.  
 1007 \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.20317>.  
 1008 [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.  
 1009 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>.  
 1010 [16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Naacl-hlt 2019. page 4171–4186, Minneapolis, Minnesota, 06 2019. Association  
 1011 for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL <https://aclanthology.org/N19-1423>.  
 1012 [17] Douglas W. Dockery, Joel Schwartz, and John D. Spengler. Air pollution and daily mortality: Associations with particulates and acid aerosols.  
 1013 *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>.  
 1014 [18] Marian Dörk, Nathalie Henry Riche, Gonzalo Ramos, and Susan Dumais. Pivotpaths: Strolling through faceted information spaces. *IEEE Transactions  
 1015 on Visualization and Computer Graphics*, 18(12):2709–2718, 12 2012. doi: 10.1109/TVCG.2012.252. URL <https://ieeexplore.ieee.org/document/6327277>.  
 1016 [19] Saeed Farzi, Soumitra Ghosh, Alberto Lavelli, and Bernardo Magnini. Get the best out of 1b llms: Insights from information extraction on clinical  
 1017 documents. page 266–276, Bangkok, Thailand, 08 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.bionlp-1.21. URL  
 1018 <https://aclanthology.org/2024.bionlp-1.21/>.  
 1019 [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  
 1020 Research Society.  
 1021 [21] Elliot Gould, Hannah S. Fraser, Timothy H. Parker, Shinichi Nakagawa, Simon C. Griffith, Peter A. Veski, Fiona Fidler, Daniel G. Hamilton, Robin N.  
 1022 Abbey-Lee, Jessica K. Abbott, Luis A. Aguirre, Carles Alcaraz, Irith Aloni, Drew Altschul, Kunal Arekar, Jeff W. Atkins, Joe Atkinson, Christopher M.  
 1023 Baker, Meghan Barrett, Kristian Bell, Suleiman Kehinde Bello, Iván Beltrán, Bernd J. Berauer, Michael Grant Bertram, Peter D. Billman, Charlie K.  
 1024 Blake, Shannon Blake, Louis Blaard, Andrea Bonisoli-Alquati, Timothée Bonnet, Camille Nina Marion Bordes, Aneesh P. H. Bose, Thomas Botterill-  
 1025 James, Melissa Anna Boyd, Sarah A. Boyle, Tom Bradfer-Lawrence, Jennifer Bradham, Jack A. Brand, Martin I. Brengdahl, Martin Bulla, Luc Bussière,  
 1026 Ettore Camerlenghi, Sara E. Campbell, Leonardo L. F. Campos, Anthony Caravaggi, Pedro Cardoso, Charles J. W. Carroll, Therese A. Catanach,  
 1027 Xuan Chen, Heung Ying Janet Chik, Emily Sarah Choy, Alec Philip Christie, Angela Chuang, Amanda J. Chunco, Bethany L. Clark, Andrea Contina,  
 1028 Garth A. Covernton, Murray P. Cox, Kimberly A. Cressman, Marco Crotti, Connor Davidson Crouch, Pietro B. D'Amelio, Alexandra Allison  
 1029 de Sousa, Timm Fabian Döbert, Ralph Dobler, Adam J. Dobson, Tim S. Doherty, Szymon Marian Drobniak, Alexandra Grace Duffy, Alison B. Duncan,  
 1030 Robert P. Dunn, Jamie Dunning, Trishna Dutta, Luke Eberhart-Hertel, Jared Alan Elmore, Mahmoud Medhat Elsherif, Holly M. English, David C.  
 1031 Ensminger, Ulrich Rainer Ernst, Stephen M. Ferguson, Esteban Fernandez-Juricic, Thalita Ferreira-Arruda, John Fieberg, Elizabeth A. Finch, Evan A.  
 1032 Fiorenza, David N. Fisher, Amélie Fontaine, Wolfgang Forstmeier, Yoan Fourcade, Graham S. Frank, Cathryn A. Freund, Eduardo Fuentes-Lillo,  
 1033 Sara L. Gandy, Dustin G. Gannon, Ana I. García-Cervigón, Alexis C. Garretson, Xuezhen Ge, William L. Geary, Charly Géron, Marc Gilles, Antje  
 1034 Girndt, Daniel Gliksman, Harrison B. Goldspiel, Dylan G. E. Gomes, Megan Kate Good, Sarah C. Goslee, J. Stephen Gosnell, Eliza M. Grames, Paolo  
 1035 Gratton, Nicholas M. Grebe, Skye M. Greenler, Maaike Griffioen, Daniel M. Griffith, Frances J. Griffith, Jake J. Grossman, Ali Güncan, Stef Haesen,  
 1036 James G. Hagan, Heather A. Hager, Jonathan Philo Harris, Natasha Dean Harrison, Sarah Syedia Hasnain, Justin Chase Havird, Andrew J. Heaton,  
 1037 María Laura Herrera-Chaustre, Tanner J. Howard, Bin-Yan Hsu, Fabiola Iannarilli, Esperanza C. Iranzo, Erik N. K. Iverson, Saheed Olade Jimoh,  
 1038 Douglas H. Johnson, Martin Johnsson, Jesse Jorna, Tommaso Jucker, Martin Jung, Ineta Kačergytė, Oliver Kaltz, Alison Ke, Clint D. Kelly, Katharine  
 1039 Keegan, Friedrich Wolfgang Keppeler, Alexander K. Killion, Dongmin Kim, David P. Kochan, Peter Korsten, Shan Kothari, Jonas Kuppler, Jillian M.  
 1040 Manuscript submitted to ACM

- 1041 Kusch, Małgorzata Lagisz, Kristen Marianne Lalla, Daniel J. Larkin, Courtney L. Larson, Katherine S. Lauck, M. Elise Lauterbur, Alan Law, Don-Jean  
 1042 Léandri-Breton, Jonas J. Lembrechts, Kiara L'Herpiniere, Eva J. P. Lievens, Daniela Oliveira de Lima, Shane Lindsay, Martin Luquet, Ross MacLeod,  
 1043 Kirsty H. Macphie, Kit Magellan, Magdalena M. Mair, Lisa E. Malm, Stefano Mammola, Caitlin P. Mandeville, Michael Manhart, Laura Milena  
 1044 Manrique-Garzon, Elina Mäntylä, Philippe Marchand, Benjamin Michael Marshall, Charles A. Martin, Dominic Andreas Martin, Jake Mitchell  
 1045 Martin, April Robin Martinig, Erin S. McCallum, Mark McCauley, Sabrina M. McNew, Scott J. Meiners, Thomas Merkling, Marcus Michelangeli,  
 1046 Maria Moiron, Bruno Moreira, Jennifer Mortensen, Benjamin Mos, Tacfeek Olatunbosun Muraina, Penelope Wrenn Murphy, Luca Nelli, Petri  
 1047 Niemelä, Josh Nightingale, Gustav Nilsonne, Sergio Nolazco, Sabine S. Nooten, Jessie Lanterman Novotny, Agnes Birgitta Olin, Chris L. Organ,  
 1048 Kate L. Ostevik, Facundo Xavier Palacio, Matthieu Paquet, Darren James Parker, David J. Pascall, Valerie J. Pasquarella, John Harold Paterson, Ana  
 1049 Payo-Payo, Karen Marie Pedersen, Grégoire Perez, Kayla I. Perry, Patrice Pottier, Michael J. Proulx, Raphaël Proulx, Jessica L. Pruitt, Veronarindra  
 1050 Ramananjato, Finaritra Tolotra Randimbiarison, Onja H. Razafindratsima, Diana J. Remnison, Federico Riva, Sepand Riyahi, Michael James Roast,  
 1051 Felipe Pereira Rocha, Dominique G. Roche, Cristian Román-Palacios, Michael S. Rosenberg, Jessica Ross, Freya E. Rowland, Deusdedith Rugemalila,  
 1052 Avery L. Russell, Suvi Ruuskanen, Patrick Saccone, Asaf Sadeh, Stephen M. Salazar, Kris Sales, Pablo Salomón, Alfredo Sánchez-Tójár, Leticia Pereira  
 1053 Santos, Francesca Santostefano, Hayden T. Schilling, Marcus Schmidt, Tim Schmoll, Adam C. Schneider, Allie E. Schrock, Julia Schroeder, Nicolas  
 1054 Schtickzelle, Nick L. Schultz, Drew A. Scott, Michael Peter Scroggie, Julie Teresa Shapiro, Nitika Sharma, Caroline L. Shearer, Diego Simón, Michael I.  
 1055 Sitvarin, Fabrício Luiz Skupien, Heather Lea Slinn, Grania Polly Smith, Jeremy A. Smith, Rahel Sollmann, Kaitlin Stack Whitney, Shannon Michael  
 1056 Still, Erica F. Stuber, Guy F. Sutton, Ben Swallow, Conor Claverie Taff, Elina Takola, Andrew J. Tanentzap, Rocío Tarjuelo, Richard J. Telford,  
 1057 Christopher J. Thawley, Hugo Thierry, Jacqueline Thomson, Svenja Tidau, Emily M. Tompkins, Claire Marie Tortorelli, Andrew Trlica, Biz R.  
 1058 Turnell, Lara Urban, Stijn Van de Vondel, Jessica Eva Megan van der Wal, Jens Van Eeckhoven, Francis van Oordt, K. Michelle Vanderwel, Mark C.  
 1059 Vanderwel, Karen J. Vanderwolf, Juliana Vélez, Diana Carolina Vergara-Florez, Brian C. Verrelli, Marcus Vinícius Vieira, Nora Villamil, Valerio  
 1060 Vitali, Julien Vollering, Jeffrey Walker, Xanthe J. Walker, Jonathan A. Walter, Paweł Waryszak, Ryan J. Weaver, Ronja E. M. Wedegärtner, Daniel L.  
 1061 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>.
- 1062 [22] Bowen Gu, Vivian Shao, Ziqian Liao, Valentina Carducci, Santiago Romero Brufau, Jie Yang, and Rishi J. Desai. Scalable information extraction from  
 1063 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>.
- 1064 [23] Ken Gu, Eunice Jun, and Tim Althoff. Understanding and supporting debugging workflows in multiverse analysis. CHI '23, page 1–19, New York,  
 1065 NY, USA, 04 2023. Association for Computing Machinery. doi: 10.1145/3544548.3581099. URL <https://dl.acm.org/doi/10.1145/3544548.3581099>.
- 1066 [24] Sonakshi Gupta, Akhlak Mahmood, Pranav Shetty, Aishat Adeboye, and Rampi Ramprasad. Data extraction from polymer literature using large  
 1067 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.
- 1068 [25] Martin Götz, Abhraneel Sarma, and Ernest H. O’Boyle. The multiverse of universes: A tutorial to plan, execute and interpret multiverses analyses  
 1069 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>.
- 1070 [26] Karlyn K. Harrod, Prabin Bhandari, and Antonios Anastasopoulos. From text to maps: Llm-driven extraction and geotagging of epidemiological  
 1071 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>.
- 1072 [27] Florian Heimerl, Qi Han, Steffen Koch, and Thomas Ertl. Citerivers: Visual analytics of citation patterns. *IEEE Transactions on Visualization and  
 1073 Computer Graphics*, 22(1):190–199, 01 2016. doi: 10.1109/TVCG.2015.2467621. URL <https://ieeexplore.ieee.org/document/7192685/authors>.
- 1074 [28] Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, Kirk  
 1075 Roberts, and Hua Xu. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American  
 1076 Medical Informatics Association*, 31(9):1812–1820, 09 2024. doi: 10.1093/jamia/ocad259. URL <https://doi.org/10.1093/jamia/ocad259>.
- 1077 [29] Wei Huang, Jianguo Tan, Haidong Kan, Ni Zhao, Weimin Song, Guixiang Song, Guohai Chen, Lili Jiang, Cheng Jiang, Renjie Chen, and Bingheng  
 1078 Chen. Visibility, air quality and daily mortality in shanghai, china. *Science of The Total Environment*, 407(10):3295–3300, 05 2009. doi: 10.1016/j.  
 1079 scitotenv.2009.02.019. URL <https://linkinghub.elsevier.com/retrieve/pii/S004896970900165X>.
- 1080 [30] Nick Huntington-Klein, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pralhad Burli, Naibin Chen, Paul Grieco, Godwin  
 1081 Ekpe, Todd Pugatch, Martin Saavedra, and Yaniv Stopnitzky. The influence of hidden researcher decisions in applied microeconomics.  
 1082 *Economic Inquiry*, 59(3):944–960, 2021. doi: 10.1111/ecin.12992. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/ecin.12992>. \_eprint:  
 1083 <https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecin.12992>.
- 1084 [31] Petra Isenberg, Tobias Isenberg, Michael Sedlmair, Jian Chen, and Torsten Möller. Visualization as seen through its research paper keywords. *IEEE  
 1085 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](https://ieeexplore.ieee.org/<br/>
  1086 document/7539364).
- 1087 [32] Eunice Jun, Maureen Daum, Jared Roesch, Sarah Chasins, Emery Berger, Rene Just, and Katharina Reinecke. Tea: A high-level language and runtime  
 1088 system for automating statistical analysis. UIST '19, page 591–603, New York, NY, USA, 10 2019. Association for Computing Machinery. doi:  
 1089 10.1145/3332165.3347940. URL <https://dl.acm.org/doi/10.1145/3332165.3347940>.
- 1090 [33] Eunice Jun, Melissa Birchfield, Nicole De Moura, Jeffrey Heer, and René Just. Hypothesis formalization: Empirical findings, software limitations,  
 1091 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](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] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.
- [45] 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>.
- [46] Ni Li, Shorouq Zahra, Mariana Brito, Clara 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>.
- [47] 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/>.
- [48] 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>.
- [49] 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/>.
- [50] 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.
- [51] Elena López-Villarrubia, Ferran Ballester, Carmen Iñíguez, 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.
- [52] 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.

- 1145 [53] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality.  
1146 volume 26. Curran Associates, Inc., 2013. URL <https://papers.nips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html>.
- 1147 [54] Suresh H. Moolgavkar. Air pollution and hospital admissions for diseases of the circulatory system in three u.s. metropolitan areas. *Journal of the Air &*  
1148 *Waste Management Association*, 50(7):1199–1206, 07 2000. doi: 10.1080/10473289.2000.10464162. URL <https://doi.org/10.1080/10473289.2000.10464162>.  
1149 Publisher: Taylor & Francis.
- 1150 [55] 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 &  
1151 Francis.
- 1152 [56] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 01 2007. doi:  
1153 10.1075/li.30.1.03nad. URL <https://www.jbe-platform.com/content/journals/10.1075/li.30.1.03nad>. Publisher: John Benjamins.
- 1154 [57] Arpit Narechania, Alireza Karduni, Ryan Wesslen, and Emily Wall. Vitality: Promoting serendipitous discovery of academic literature with  
1155 transformers & visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):486–496, 01 2022. doi: 10.1109/TVCG.2021.3114820.  
1156 URL <https://ieeexplore.ieee.org/document/9552447/>.
- 1157 [58] Bart Ostro, Rachel Broadwin, Shelley Green, Wen-Ying Feng, and Michael Lipsett. Fine particulate air pollution and mortality in nine california  
1158 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.
- 1159 [59] Roger D. Peng, Francesca Dominici, and Thomas A. Louis. Model choice in time series studies of air pollution and mortality. *Journal of the Royal*  
1160 *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>.
- 1161 [60] Maciej P. Polak and Dane Morgan. Extracting accurate materials data from research papers with conversational language models and prompt  
1162 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.
- 1163 [61] 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,  
1164 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>.
- 1165 [62] Abhraneel Sarma, Alex Kale, Michael Moon, Nathan Taback, Fanny Chevalier, Jessica Hullman, and Matthew Kay. multiverse: Multiplexing  
1166 alternative data analyses in r notebooks (version 0.6.2). *OSF Preprints*, 2021. URL <https://github.com/MUCollective/multiverse>.
- 1167 [63] Marko Sarstedt, Susanne J. Adler, Christian M. Ringle, Gyeongcheol Cho, Adamantios Diamantopoulos, Heungsun Hwang, and Benjamin D.  
1168 Lienggaard. 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>.  
1169 \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/jpim.12738>.
- 1170 [64] Mara Schilling-Wilhelmi, Martín Ríos-García, Sherjeel Shabih, María Victoria Gil, Santiago Miret, Christoph T. Koch, José A. Márquez, and Kevin  
1171 Maik Jablonka. From text to insight: large language models for chemical data extraction. *Chemical Society Reviews*, 54(3):1125–1150, 2025. doi:  
1172 10.1039/D4CS00913D. URL <https://pubs.rsc.org/en/content/articlelanding/2025/cs/d4cs00913d>. Publisher: Royal Society of Chemistry.
- 1173 [65] 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.
- 1174 [66] Veronica Sciannameo, Daniele Jahier Pagliari, Sara Urru, Piercesare Grimaldi, Honoria Ocagli, Sara Ahsani-Nasab, Rosanna Irene Comoretto, Dario  
1175 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>.
- 1176 [67] R. Silberzahn, E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtry, Š. Bahník, F. Bai, C. Bannard, E. Bonnier, R. Carlsson, F. Cheung,  
1177 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,  
1178 T. J. Heaton, K. Hederos, M. Heene, A. J. Hofelich Mohr, F. Höglund, K. Hui, M. Johannesson, J. Kalodimos, E. Kaszubowski, D. M. Kennedy, R. Lei,  
1179 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,  
1180 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,  
1181 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  
1182 transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, 1(3):337–356, 09 2018. doi:  
1183 10.1177/2515245917747646. URL <https://doi.org/10.1177/2515245917747646>. Publisher: SAGE Publications Inc.
- 1184 [68] Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows  
1185 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.
- 1186 [69] Jan Simson, Fiona Draxler, Samuel Mehr, and Christoph Kern. Preventing harmful data practices by using participatory input to navigate the  
1187 machine learning multiverse. CHI '25, page 1–30, New York, NY, USA, 04 2025. Association for Computing Machinery. doi: 10.1145/3706598.3713482.  
1188 URL <https://dl.acm.org/doi/10.1145/3706598.3713482>.
- 1189 [70] Imad Tbahriti, Christine Chichester, Frédérique Lisacek, and Patrick Ruch. Using argumentation to retrieve articles with similar citations: An  
1190 inquiry into improving related articles search in the medline digital library. *International Journal of Medical Informatics*, 75(6):488–495, 06 2006. doi:  
1191 10.1016/j.ijmedinf.2005.06.007. URL <https://www.sciencedirect.com/science/article/pii/S1386505605000894>.

- 1197 [71] G. Touloumi, E. Samoli, M. Pipikou, A. Le Tertre, R. Atkinson, and K. Katsouyanni. Seasonal confounding in air pollution and health time-series  
 1198 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>.
- 1199  
 1200 [72] Kayo Ueda, Nitta ,Hiroshi ,Ono ,Masaji , , and Ayano Takeuchi. Estimating mortality effects of fine particulate matter in japan: A comparison of time-  
 1201 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.  
 1202 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>.
- 1203 [73] 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  
 1204 freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, 7, 11 2016.  
 1205 doi: 10.3389/fpsyg.2016.01832. URL <https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2016.01832/full>. Publisher: Frontiers.
- 1206 [74] 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>.
- 1207 [75] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes,  
 1208 Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson,  
 1209 Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal  
 1210 of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.
- 1211 [76] Hadley Wickham, Joe Cheng, and Aaron Jacobs. *ellmer: Chat with Large Language Models*, 2025. URL <https://CRAN.R-project.org/package=ellmer>.  
 1212 R package version 0.1.1.
- 1213 [77] Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. Large language  
 1214 models for generative information extraction: A survey. doi: 10.48550/arXiv.2312.17617.
- 1215 [78] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for  
 1216 language understanding. doi: 10.48550/arXiv.1906.08237.
- 1217  
 1218  
 1219  
 1220  
 1221  
 1222  
 1223  
 1224  
 1225  
 1226  
 1227  
 1228  
 1229  
 1230  
 1231  
 1232  
 1233  
 1234  
 1235  
 1236  
 1237  
 1238  
 1239  
 1240  
 1241  
 1242  
 1243  
 1244  
 1245  
 1246  
 1247  
 1248 Manuscript submitted to ACM