

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: • **Applied computing** → *Document analysis*; • **Human-centered computing** → *Empirical studies in HCI*.

14     Additional Key Words and Phrases: Large language models

15     **ACM Reference Format:**

16     Anonymous Author(s). 2025. An LLM-based pipeline for understanding decision choices in data analysis from published literature.  
17     In *Proceedings of CHI Conference on Human Factors in Computing Systems (CHI'26)*. ACM, New York, NY, USA, 22 pages. <https://doi.org/XXXXXX.XXXXXX>

18     **1 Introduction**

19     TODO: need references

20     Decisions are made at every stage of data analysis, from initial data collection and preprocessing to modeling. One  
21     might expect well-trained researchers to make similar choices when faced with the same analytical task, yet evidence  
22     suggests otherwise. Many-analyst experiments show that independent analysts often arrive at markedly different  
23     conclusions, even when analyzing the same dataset to answer the same research question [8, 20, 47]. This variation in  
24     analytical decision-making, described by Gelman and Loken [19] as the “garden of forking paths,” can undermine the  
25     quality and credibility of reported results and hinder comparability across studies. For junior researchers who lack  
26     guidance, this variability may lead to over reliance on default statistical software settings or arbitrary choices made  
27     without clear justification.

28     A common approach to investigate uncertainty in decision choices is sensitivity analysis, where researchers systematically  
29     vary key decisions in their analysis to assess the robustness of their findings. Multiverse analysis extends this  
30     idea by evaluating *all* plausible combinations of analytical choices to examine how results vary across the full decision  
31     space [7, 44]. However, what an analyst consider “reasonable” is subjective and may not reflect the full range of options  
32     commonly used in practice. Even when a reasonable set of alternatives is tested, the sensitivity analysis may be of

---

33     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  
34     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  
35     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  
36     redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

37     © 2025 ACM.

38     Manuscript submitted to ACM

39     Manuscript submitted to ACM

53 limited interest to other researchers facing a similar problem, who are seeking evidence to inform comparable decision  
54 choices and their rationale. Ideally, decision-making in applied research would be studied by following experienced  
55 analysts throughout the entire analysis process to capture their reasoning. In reality, this is rarely feasible and not  
56 scalable.  
57

58 While individual studies may not capture the full range of reasonable decision options, crowdsourcing decisions  
59 from a collection of studies on a shared theme creates a “many-analyst” setting that reveals how analysts make choices  
60 and justify them in practice. Classic research training typically involves reading through the literature to understand  
61 how decisions are made and to learn the common choices. This process now has the possibility to be automated at scale  
62 given recent LLMs’ ability to follow instructions to extract structured information from unstructured text. In this work,  
63 we propose a new approach for studying data analysis decision choices by automatically extracting decisions from  
64 scientific literature using Large Language Models (LLMs). We develop a tabular schema to record decisions, automate  
65 the extraction process with LLMs, and introduce a new paper similarity measure based on decision similarity, which  
66 serves as a distance metric for dimension reduction methods to visualize papers group according to their decision  
67 patterns.  
68

69 We apply this workflow to a set of 56 air pollution modelling studies estimating the effect of particulate matter (PM<sub>2.5</sub>  
70 or PM<sub>10</sub>) on mortality and hospital admissions. This type of studies is typically analyzed using Poisson generalized linear  
71 models (GLMs) or generalized additive models (GAMs). Analysis of the extracted decisions reveals common choices for  
72 decisions considered in this type of studies such as the number of knots or degree of freedom for smoothing methods  
73 and the temporal lags for time and weather variables. Multi-dimensional scaling on the paper similarity distance finds  
74 three distinct clusters corresponding to different smoothing methods – LOESS, natural spline, and smoothing spline –  
75 used in European and U.S. studies. These findings align with the APHENA project [29], which synthesizes research  
76 from multiple studies in Europe and North America by expert investigators.  
77

78 In this workflow, we also provide detailed documentation on the validation and standardization of LLM outputs.  
79 Because LLMs generate results probabilistically, it is not yet clear how these outputs should be validated for downstream  
80 analysis in practice. We outline the validation and standardization process, including the use of a developed Shiny  
81 application in R for reviewing decisions, the types of edits made through validation, and secondary standardization of  
82 decisions. Additionally, we conduct sensitivity across different LLM providers and assess the reproducibility of the  
83 text extraction from single LLM models. We aim to offer guidance for future studies seeking to extract structured  
84 information from unstructured text using LLMs.  
85

86 In summary, the contribution of this work includes:  
87

- 88 • A scalable and automated approach to study data analysis decision choices through extracting of decisions from  
89 published scientific literature using LLMs,
- 90 • A new method to construct paper similarities based on the decision and the semantic similarity of their rationale,  
91 • A data schema for summarizing decisions made in applied studies in a tidy format,  
92 • A shiny GUI tool for validating LLM outputs for editing tables, and  
93 • A dataset of decisions and rationale, along with metadata, compiled from 62 studies in air pollution mortality  
94 modelling.  
95

**105 2 Related work****106 2.1 Decision-making in data analysis**

108 Data analysis involves making choices at every step, from initial data collection, data pre-processing to model specification,  
109 and post-processing. Each decision represents a branching point where analysts choose a specific path to follow,  
110 and the vast number of possible choices analysts can take forms what Gelman and Loken [19] describe as the “garden  
111 of forking paths”. While researchers may hope their inferential results are robust to the specific path taken through  
112 the garden, in practice, different choices can lead to substantially different conclusions. This has been empirically  
113 demonstrated through “many analyst experiments”, where independent research groups analyze the same dataset to  
114 address the same research questions with their own chosen analytic approach. A classic example is Silberzahn et al.  
115 [47], where researchers reported an odds ratio from 0.89 to 2.93 for the effect of soccer players’ skin tone on the number  
116 of red cards awarded by referees. Similar variability has been observed in structural equation modeling [45], applied  
117 microeconomics [23], neuroimaging [8], and ecology and evolutionary biology [20].  
118

119 Examples like above have rendered decision-making in data analysis as a subject to study in human computer  
120 interaction. To understand how analysts making decisions during data analysis and navigating the garden of forking  
121 path, researchers have conducted qualitative interviews with analysts on data analysis practices [2, 25, 32]. Visualization  
122 tools have also been explored to communicate the decision process through analytic decision graphics (ADG) [33]. In  
123 fairness machine learning literature, Simson et al. [48] contributed a reusable workflow that supports participatory input  
124 to democratize decisions in machine learning algorithms related to fairness, privacy, interpretability and performance.  
125 Conducting qualitative studies through interviews to study how assumptions and decisions are made in data analysis  
126 practices takes a significant amount of time and effort, and the findings may not generalize to other contexts. While  
127 published research papers may not provide a complete picture of the decision-making process, they do contain valuable  
128 information about the choices made by analysts and the rationale behind them. With recent advances in Large Language  
129 Models (LLMs), it has become possible to automatically extract structured information from unstructured text. This  
130 could provide a scalable way to study decision-making practices in data analysis.  
131

**132 2.2 Uncertainty in data analysis**

133 On top of qualitative studies, software tools have also developed to incorporate potential alternatives in the analysis  
134 workflow. The DeclareDesign package [7] introduces the MIDA framework for researchers to declare, diagnose, and  
135 redesign their analyses to produce a distribution of the statistic of interest, which has been applied in the randomized  
136 controlled trial study [6]. The multiverse package [34, 44] provides a framework for researchers to conduct multiverse  
137 analysis to systematically explore how different choices affect results and to report the range of plausible outcomes that  
138 arise from alternative analytic paths.  
139

**140 2.3 Automatic information extraction with LLMs****141 2.4 Visualization on scientific literature**

142 With the growing volume of scientific publications and the difficulty of navigating the literature to stay informed,  
143 there is increasing interest in developing tools to visualize and recommend scientific papers. These systems link papers  
144 based on their similarity and relevance, typically determined by keywords [24], citation information (e.g. citation list,  
145 co-citation) [14], or combinations with other relevant paper metadata (e.g. author, title) [5, 15, 18, 21]. Recent approaches  
146 incorporate text-based information using topic modelling [1], argumentation-based information retrieval [49], and  
147

157 text embedding [40]. While metadata and high-level text-based information are useful for finding relevant papers,  
 158 researchers also need tools that help them *make sense* of the literature rather than simply *locating* it. In applied data  
 159 analysis, one interest is to understand how studies differ or align in their analytical approaches. Capturing the decisions  
 160 and reasoning expressed in analyses on a shared theme enables the calculation of similarity metrics based on these  
 161 choice and their underlying rationale, which supports clustering and visualizing paper to identify common practices in  
 162 the field.  
 163

### 165 3 Methods

166 In this section, we present the workflow for extracting decisions from published literature using Large Language Models  
 167 (LLMs). We first describe the data structure for recording decisions, followed by the four main steps: 1) automatic  
 168 extraction from literature with LLMs, 2) validation and standardization of LLM outputs, 3) calculation of paper similarity,  
 169 and 4) visualization paper similarity using clustering or dimension reduction methods. The section concludes with an  
 170 illustration summarizing the workflow.  
 171

#### 172 3.1 Record decisions in data analysis

173 In the study of the health effects of outdoor air pollution, one area of interest is the association between short-term,  
 174 day-to-day changes in particulate matter air pollution and daily mortality counts. This question has been studied  
 175 extensively by researchers across the globe and in the US, it serves to provide scientific evidence for to guide public policy  
 176 on setting the National Ambient Air Quality Standards (NAAQS) for air pollutants. While individual modelling choices  
 177 vary, these studies often share a common structure: they adjust for meteorological covariates such as temperature and  
 178 humidity, apply temporal or spatial treatments, like including lagged variables and may estimate the effect by city or  
 179 region before combining results. This naturally forms a “many-analyst” experiment setting where different researchers  
 180 analyze similar data to address the same scientific question and the analyses are documented in published papers.  
 181

182 Consider the following excerpt from Ostro et al. [41] that describes the modelling approach to provide evidence of  
 183 an association between daily counts of mortality and ambient particulate matter (PM10):  
 184

185     Based on previous findings reported in the literature (e.g., Samet et al. 2000), the basic model included a  
 186     smoothing spline for time with 7 degrees of freedom (df) per year of data. This number of degrees of  
 187     freedom controls well for seasonal patterns in mortality and reduces and often eliminates autocorrelation.  
 188

189 This sentence encode the following components of a decision:

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

196 To record these decisions in a tabular format, we follow the tidy data principle [52], which states each variable should  
 197 be in a column and each observation in a row. For our purpose, each row represents a decision made by the authors  
 198 in a paper and an analysis often include multiple decisions. To retain the original context of the decision, we extract  
 199 the original text in the paper, without paraphrase or summarization. The decision choice above is a parameter choice  
 200 of a statistical method applied to the variable. Analyses also include other types of decisions, such as temporal and  
 201 spatial treatments, for example, the choice of lagged exposure for certain variables or whether the model is estimated  
 202

209 collectively or separated for individual locations. These decisions don't have a specific method or parameter, but should  
 210 still be recorded with the variable, type (spatial or temporal), reason, and decision fields.  
 211

212 Given the writing style and the quality of the analysis itself, multiple decisions may be combined in one sentence  
 213 and certain fields, e.g. decision and reason, may be omitted. Consider the following excerpt from Ostro et al. [41]:  
 214

215 Other covariates, such as day of the week and smoothing splines of 1-day lags of average temperature  
 216 and humidity (each with 3 df), were also included in the model because they may be associated with  
 217 daily mortality and are likely to vary over time in concert with air pollution levels.  
 218

219 This sentence contains four decisions: two for temperature (the temporal lag and the smoothing spline parameter)  
 220 and two for humidity and should be structured as separate entries:  
 221

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

234  
 235 Notice in the example above, the reason field are recorded as NA. This is because the stated rationale ("and are  
 236 likely to vary over time in concert with air pollution levels") only supports the general inclusion of temporal lags but  
 237 does not justify the specific choice of 1-day lag over other alternatives, for example, 2-day average of lags 0 and 1 and  
 238 single-day lag of 2 days. Similar scenario can happen when a direct decision is missing while a reason is provided  
 239 ("done by minimizing Akaike's information criterion"), as in Katsouyanni et al. [28]:  
 240

241 The inclusion of lagged weather variables and the choice of smoothing parameters for all of the weather  
 242 variables were done by minimizing Akaike's information criterion.  
 243

### 244 3.2 Extract decisions automatically from literature with LLMs

245 Manually extracting decisions from published papers is labor-intensive and time-consuming. With Large Language  
 246 Models (LLMs), it has become possible to automatically extract structured information from unstructured text by  
 247 supplying a set of PDF documents and a prompt for instruction. Text recognition from PDF document relies on Optical  
 248 Character Recognition (OCR) to convert scanned images into machine-readable text – capability currently offered by  
 249 Anthropic Claude and Google Gemini. In the prompt, we assign the LLM a role as an applied statistician and instruct it  
 250 to generate a markdown file containing a JSON block that extract decisions from the PDF in the format described in  
 251 Section 3.1. We also provide a set of instructions and examples on the potential missing of reason and decision fields.  
 252 Prompt engineering techniques [13, 55] are used to optimize the prompt script. The full prompt feed to the LLM is  
 253 provided in the Appendix. We use the `chat_PROVIDER()` functions from the `ellmer` package [54] in R to obtain the  
 254 output with Gemini and Claude API.  
 255

### 261    3.3 Validate and standardize LLM outputs

262  
 263    The LLM outputs need to be validated and standardized before further analysis. Validation focuses on ensuring the  
 264    correctness of the extracted decisions by LLMs, while standardization aims to ensure consistency in variable and model  
 265    names across papers, given authors may express the same concept in different ways. For example, “mean temperature”,  
 266    “average temperature”, and “temperature” all refer to the same variable, which can be all standardized to “temperature”  
 267    for consistency. To help with the validation and standardization process, we developed a Shiny application that provides  
 268    an interactive interface for users to review and edit the LLM outputs. A Shiny application takes a CSV of extracted  
 269    decisions as input and allows three types of edits: 1) *overwrite* – modify the content of a particular cell, 2) *delete* –  
 270    remove a particular irrelevant decision, and 3) *add* – manually enter a missing decision. Figure 1 illustrates the *overwrite*  
 271    action for standardizing the variable NCtot (The number concentration of urban background particles <100 nm in  
 272    diameter) to “pollution”: the user enters a predicate function in the filter condition box on the left panel, and the filtered  
 273    data will appear interactively in the right panel. The user can then specify the variable to overwrite and the new value  
 274    and the corresponding cells in the right panel will be updated. This change need to be confirmed by pressing the “Apply  
 275    changes” button to update the full dataset. The corresponding tidyverse [53] code will then be generated in the left  
 276    panel to be included in an R script, and the edited table can be downloaded for future analysis.  
 277  
 278  
 279

### 281    3.4 Calculate paper similarity and visualization

282  
 283    Once the output has been extracted and validated, the decisions can be treated as data for further analysis. In this  
 284    section, we construct a distance metric between pairs of papers based on the similarity of their decision choices. This  
 285    metric can then be used as a distance matrix among papers for clustering, dimension reduction, and visualization.  
 286

287    For each paper pair, a decision is considered comparable if the papers share the same variable and decision type, for  
 288    example, a parameter decision on temperature or the temporal decision on humidity. For two decisions to be considered  
 289    similar, both the decision choice and the rationale are taken into account. A similar choice indicates a similar final  
 290    decisions are made in the analysis, whereas a similar reason reflects a shared rationale or justification for the choice,  
 291    even when the choices themselves differ, potentially due to differences in the underlying data. To assign numerical  
 292    value for measuring the similarity, we use the semantic similarity from text model, using the text package [30]. We  
 293    first obtain the text embedding for all the reason and decisions and calculate the cosine similarity between the matched  
 294    reason and decisions. For parameter type decisions, the statistical method used also contributes to the similarity of the  
 295    decision. Since semantic similarity cannot fully capture the difference betweenit statistical methods (the difference  
 296    between smoothing spline and natural spline is not well represented by the textual difference of “smoothing” and  
 297    “natural”), method similarity is encoded as binary: 1 if the two papers used the same method, and 0 otherwise. The  
 298    paper similarity is then computed as the average similarity across all the matched methods, decisions, and reasons. The  
 299    resulting paper similarity metric can be interpreted as a distance measure to cluster and visualize papers based on their  
 300    decision choices.  
 301  
 302

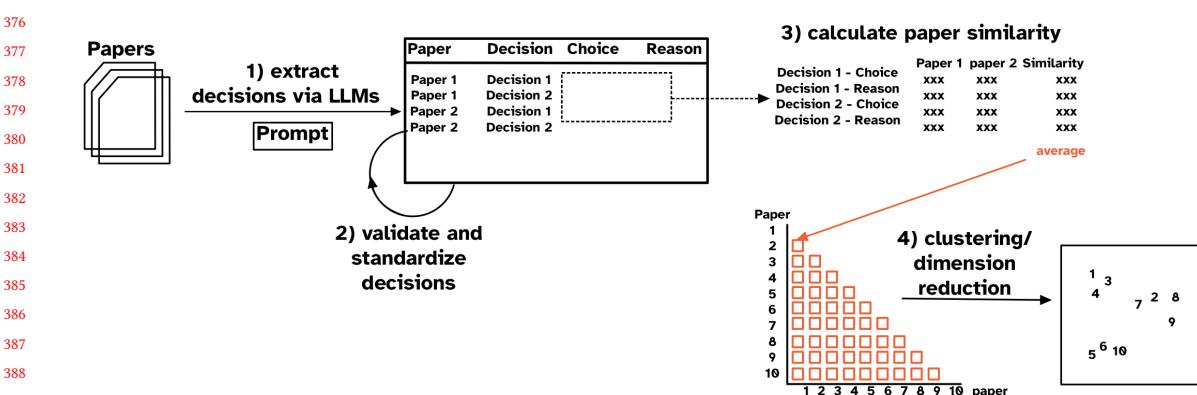
303    Because analyses vary in the decisions they report, the number of matched decisions differs across paper pairs. In  
 304    practice, some studies may not fully report the decision and reason for every choice made, leading to missing data for  
 305    the matched decisions. Although paper similarity can be calculated based on all available matched decisions, cares  
 306    should be taken for pairs with only a small number of matches, as the paper similarity may be overly influenced by one  
 307    or two decisions. To address this, users may focus on a set of decisions shared across papers and on papers that report a  
 308    minimal number of these decisions when calculating paper similarity.  
 309  
 310  
 311

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, replace the variable NCtot with “pollutant”. (3) After clicking the Confirm button, the corresponding tidyverse code is generated, and the table view returns to its original unfiltered view with the edits applied. The edited data can be downloaded by clicking the Download CSV button.

365 **3.5 Summary**

366  
 367 Figure 2 summarises the entire workflow for extracting decisions from published literature using Large Language  
 368 Models (LLMs) and analyzing the extracted decisions. Once researchers have identified a set of literature of interest and  
 369 a prompt to instruct LLMs to extract decisions from the literature. The outputs from LLM need to be validated and  
 370 standardized before further analysis due to authors' varied writing styles. The validated data can then be used to conduct  
 371 exploratory analysis of decision choices and one task is to calculate paper similarity based on the decision similarity.  
 372 This paper similarity measure can be seen as a distance metric among papers, which can be used for clustering and  
 373 dimension reduction for visualizing papers.  
 374



390 Fig. 2. The workflow for extracting decisions from published literature using Large Language Models (LLMs) and analyzing the  
 391 extracted decisions. The workflow consists of four main steps: (1) Extract decisions automatically from literature with LLMs, (2)  
 392 Validate and standardize LLM outputs, (3) Calculate paper similarity and visualization, and (4) visualization with clustering or  
 393 dimension reduction methods.  
 394

395 **4 Results**

396 From the 56 studies examining the effect of particulate matters ( $PM_{10}$  and  $PM_{2.5}$ ) on mortality and hospital admission,  
 397 we focus on the baseline model reported in each paper, excluding secondary models (e.g. lag-distributed models) and  
 398 sensitivity analysis. We also exclude decisions on other pollutants, such as nitrogen dioxide ( $NO_2$ ). This yields 242  
 399 decisions extracted using Gemini, averaging approximately 4 decisions per paper.  
 400

401 **4.1 Validation and standardization of LLM outputs**

402 Table 2. Summary of validation and standardization edits made during the review process.  
 403

404 Reason	405 Count
406 Remove decisions out of scope: other pollutants and sensitivity analysis	50
407 Edit made to recode smoothing parameter unit to per year	45
408 Duplicates	9
409 Fix incorrect capture	9
410 Manuscript submitted to ACM	

417 Table 2. Summary of validation and standardization edits made during the review process.  
418  
419

420 Reason	421 Count
422 Edit made due to decisions are too general, e.g. minimum of 1 df per year was required	6
423 Remove decisions related to definition of variables, e.g. season	5
424 Total	124

425  
426  
427 Table 2 summarizes the number of edits made during the review process using the Shiny application. These edits  
428 fall into two main categories: 1) correcting LLM outputs and 2) standardizing extracted decision. The first category  
429 includes fixing incorrect captures, removing non-decision (e.g. definition of variables), removing duplication, excluding  
430 irrelevant decisions (e.g. sensitivity analyses), and excluding decisions whose stated reasons reflect general guidelines  
431 rather than actual choices (e.g. “minimum of 1 degree of freedom per year is required”).

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

- 435 • **temperature**: “mean temperature”, “average temperature”, “temperature”, “air temperature”, “ambient tempera-  
436 ture”
- 437 • **humidity**: “dewpoint temperature” and its hyphenated variants, relative humidity”, “humidity”
- 438 • **PM**: “pollutant”, “pollution”, “particulate matter”, “particulate”, “PM10”, “PM2.5”
- 439 • **time**: “date”, “time”, “trends”, “trend”

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

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

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

450 This variability makes manual standardization impractical, hence we apply a secondary LLM process (claude-3-  
451 7-sonnet-latest) using the ellmer package to convert temporal decisions into a consistent format: **multi-day**: lag  
452 [start]-[end] and **single-day**: lag [start], . . . , lag [end]. For instance, “6-day average” is converted to  
453 “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”.

## 461 4.2 Exploratory analysis of decision choices

462 As raised in Section 3.1, not all decisions reported in the literature include both the decision choice and the rationale.  
463 Some decisions may only report the choice without a stated reason, while others may provide a reason without  
464 specifying the exact choice made. Table 3 summarizes the missingness of the decisions and reason for the extracted  
465 decisions. While 2% of decisions are complete for both decision and reasons, 55% of decisions lack a stated rationale

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%)

477  
 478 for the choice. This reflects a common reporting practice in the field, where authors often present the decision itself  
 479 without providing a justification, e.g. “We decide to use  $x$  degree of freedom for variable  $y_1$  and  $y_2$ ”. This also includes  
 480 cases where authors provide general guidelines for selecting the parameter, but the rationale is too broad to justify the  
 481 specific choice made (hence validated as NA in Section 4.1).

482  
 483 Table 4. Count of variable-type decisions in the Gemini-extracted decisions. The most commonly reported decision are the parameter  
 484 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

500  
 501 Table 4 lists the eight most frequently reported decision: parameter and temporal choice for time, PM, temperature,  
 502 and humidity. While a wider list of variables have been used in the analysis, these four variables are most commonly  
 503 included in baseline models. Parameter choices for time, temperature, and humidity are typically made on the use of  
 504 smoothing parameter for the smoothing method (natural spline and smoothing spline), whereas temporal choices are  
 505 commonly reported for PM, temperature, and humidity for the number of lag to consider in the model.

506  
 507 Table 5. Options captured for parameter choices for time, humidity, and temperature variables in the Gemini-extracted decisions.  
 508 The choices for natural spline knots are generally less varied than the degree of freedom choices for smoothing spline. Choices for  
 509 temperature and humidity tend to be close, given they are both weather related variables, while the choices for time are more varied  
 510 inherently.

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

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

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

525  
 526 Table 5 presents the parameter-related decisions extracted for spline methods (natural and smoothing spline) applied  
 527 to variable time, humidity and temperature. These decisions concern the number of knots or degree of freedom, with all  
 528 values standardized to a *per year* scale for consistency. The selection of knot for natural spline has less variation than  
 529 the degree of freedom choices for smoothing spline. Choices for temperature and humidity are generally similar, given  
 530 they are both weather related variables, whereas choices for time are more varied. This tabulation provides a reference  
 531 set for common parameter choices for future studies and help to identify anomalies and special treatment in practice.  
 532 For example, the choice of 7.7 degree of freedom reported in Castillejos et al. [12] may prompt analysts to seek further  
 533 justification. By cross comparing with other reporting, some decisions appear ambiguous. For example, in Moolgavkar  
 534 [38] and Moolgavkar [39], the reported value of 30 and 100 degrees of freedom for time may be understandable for  
 535 experienced domain researcher, it could be unclear for junior analysts as to whether they apply to the full 9 year period  
 536 or on a per-year basis. We also observe a different report style from Schwartz [46], where smoothing spline parameters  
 537 are expressed as a proportion of the data (“5% of the data” and “5% of the data”) rather than fixed numerical value.  
 538

540  
 541 Table 6. Options captured for temporal lag choices for PM, temperature, and humidity variables in the Gemini-extracted decisions.  
 542 Both single-day lags and multi-day average lags are commonly used, generally considering up to five days prior (lag 5).  
 543

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

564  
 565 Similarly, Table 6 summarizes the temporal lag choices for PM, temperature, and humidity. For single-day lags, the  
 566 lags are considered up to 13 days (approximately two weeks). For multi-day averages, 3-day and 5-day averages are  
 567 most common, although other choices such as 2-4 day average are also observed as in López-Villarrubia et al. [36]:  
 568

569 In particular, lags 0 to 1 and lags 2 to 4 averages of temperature, relative humidity, and barometric  
 570 pressure were considered as meteorological variables.  
 571

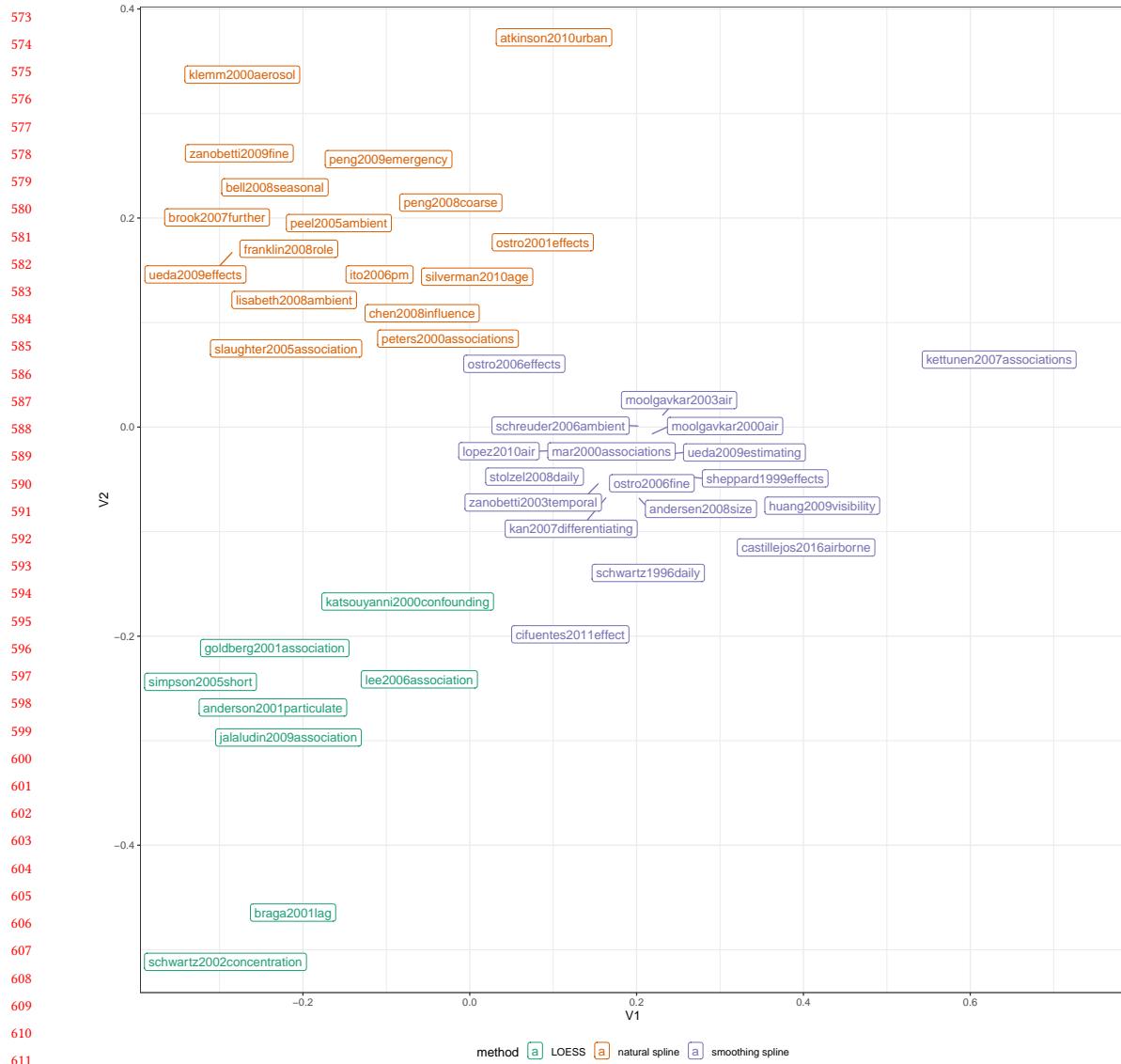


Fig. 3. The multi-dimensional scaling (MDS) based on paper similarity distance for `length(good_pp)` air pollution mortality modelling papers, colored by the smoothing method used. The MDS reveals the three distinct groups of papers, corresponds to LOESS, natural spline, and smoothing spline. These groups corresponds to the different modelling strategies debated in the European and U.S. studies, as documented in the APHENNA project [29].

### 4.3 Paper similarity and clustering

Given the number of decisions reported in Table 4, we focus on the six most common variable-type decisions for calculating paper similarity: parameter choices for time, temperature, and humidity, and temporal lag choices for PM, temperature, and humidity. We also restrict our analysis to papers that report at least three of these six decisions, resulting

Manuscript submitted to ACM

in 48 papers for the similarity analysis. This ensures that the paper similarity metric is based on a sufficient number of comparable decisions. We use the default text embedding model (BERT) in the `text` package and cosine similarity to compute the similarity score. Sensitivity analysis on different text embedding model is checked in Section 4.4.3. Paper similarity is then calculated as the average of decision similarity for each paper pair. The resulting distance matrix is then used for multi-dimensional scaling (MDS) in Figure 3. The two MDS dimension reveals three clusters correspond to the three smoothing methods used in these analyses: LOESS, natural spline, and smoothing spline. This grouping aligns with the modelling strategies seen in large-scale analysis, such as the U.S. NMMAPS study [43] and the European APHEA [27] and APHEA2 [28] project.

To reconcile these differences, the APHENA project [29] was launched with the aim to “assess the consistency across Europe and North America when estimated using a common analytic protocol and to explore possible explanations for any remaining variation”. While multi-dimensional scaling in Figure 3 shows the match of three clusters with three smoothing methods, this is not inconsistent with the APHENA project [29] that the amount of smoothing to have a more important role than the method of smoothing for estimating the effect of PM on public health variables. The similarity metric we proposed focuses on the variation of choices across analyses, without directly assessing how those choices influence results. By pooling decision choices from multiple studies with LLMs, it becomes much easier to reveal common practices and difference in research practices, highlighting decisions that require further sensitivity analyses to assess their impact. The different smoothing methods revealed in Figure 3 are consistent with the analysis by Peng et al. [42] and Tou loumi et al. [50] that compares different smoothing methods and rationale for selecting smoothing parameters.

#### 4.4 Sensitivity analysis

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

**4.4.1 LLM reproducibility.** We assess the reproducibility of Gemini’s text extraction (`gemini-2.0-flash`) by repeating the task five times for each of the 62 papers and perform pairwise comparison between runs. This generates  $5 \times 4 / 2 \times 62 = 620$  possible comparisons for both “reason” and “decisions” fields. Comparisons where the runs produced a different number of decisions were excluded, as this would require manual alignment. After filtering, 449 out of 620 (72%) remained. Table 7 prints the decisions in Andersen et al. [3] across two runs and all the four decisions are identical with no difference.

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

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

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

679

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

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

701 Table 8 summarizes the number of differences observed in each pairwise comparison. Among all comparisons, 80%  
 702 produce the identical text in reason and decision. The discrepancies come from the following two reasons: 1) Gemini  
 703 extracted different length for the same decision, e.g. in Kan et al. [26], some runs may extract “singleday lag models  
 704 underestimate the cumulative effect of pollutants on mortality 2day moving average **of current and previous day**  
 705 **concentrations** (lag=01)”, while others extract “singleday lag models underestimate the cumulative effect of pollutants  
 706 on mortality 2day moving average (lag=01)”. Similarity, for decisions, some runs yield “10 df for total mortality”, while  
 707 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  
 708 first run generates NAs in the reasons, but the remaining four runs are identical. In Ueda et al. [51] and Castillejos  
 709 et al. [12], runs 1 and 5 fail to extract the reasons and produce the same incomplete version, whereas runs 2, 3, and 4  
 710 produce accurate versions with reasons populated.

711 **4.4.2 LLM models.** Reading text from PDF document requires Optical Character Recognition (OCR) to convert images  
 712 into machine-readable text, which currently is only supported by Anthropic Claude (claude-3-7-sonnet-latest) and  
 713 Google Gemini (gemini-2.0-flash). We compare the number of decisions extracted by Claude and Gemini across all  
 714 62 papers in Figure 4. Each point represents a paper, with the x- and y-axes showing the number of decisions extracted  
 715 by Claude and Gemini, respectively. The dashed 1:1 line marks where both models extract the same number of decisions.  
 716 While both models extract decisions irrelevant to our analysis, such as sensitivity analyses and secondary analyses,  
 717 Claude’s extractions tend to include more of these irrelevant decisions, examples of these include 1) the definition  
 718 of “cold day” and “hot day” indicators in Dockery et al. [17] (“defined at the 5th/ 95th percentile”), 2) decisions relate  
 719 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. [22], and 3) the definition  
 720 of black smoke and in Katsouyanni et al. [28] for secondary analysis (“restrict to days with BS concentrations below  
 721 150  $\mu\text{g}/\text{m}^2$ ). While Gemini also capture these irrelevant decisions, such as “0-4 lag days” for air pollution exposure  
 722 Manuscript submitted to ACM

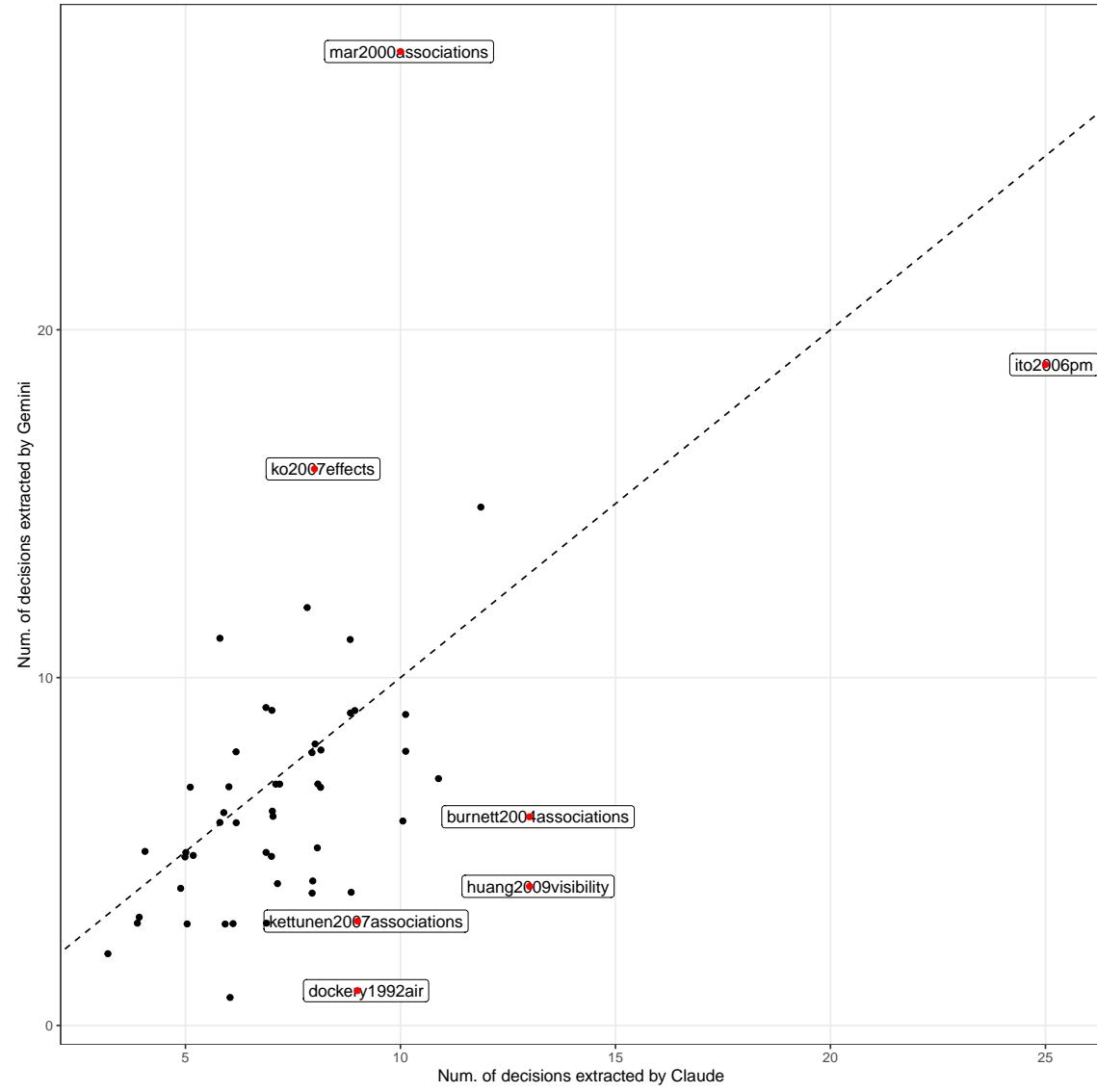


Fig. 4. Comparison of decisions extracted by Claude and Gemini. Each point represents a paper, with the x- and y-axes showing the number of decisions extracted by Claude and Gemini, respectively. The dashed 1:1 line marks where both models extract the same number of decisions. More points fall below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

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. [37]. 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. [28]

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.

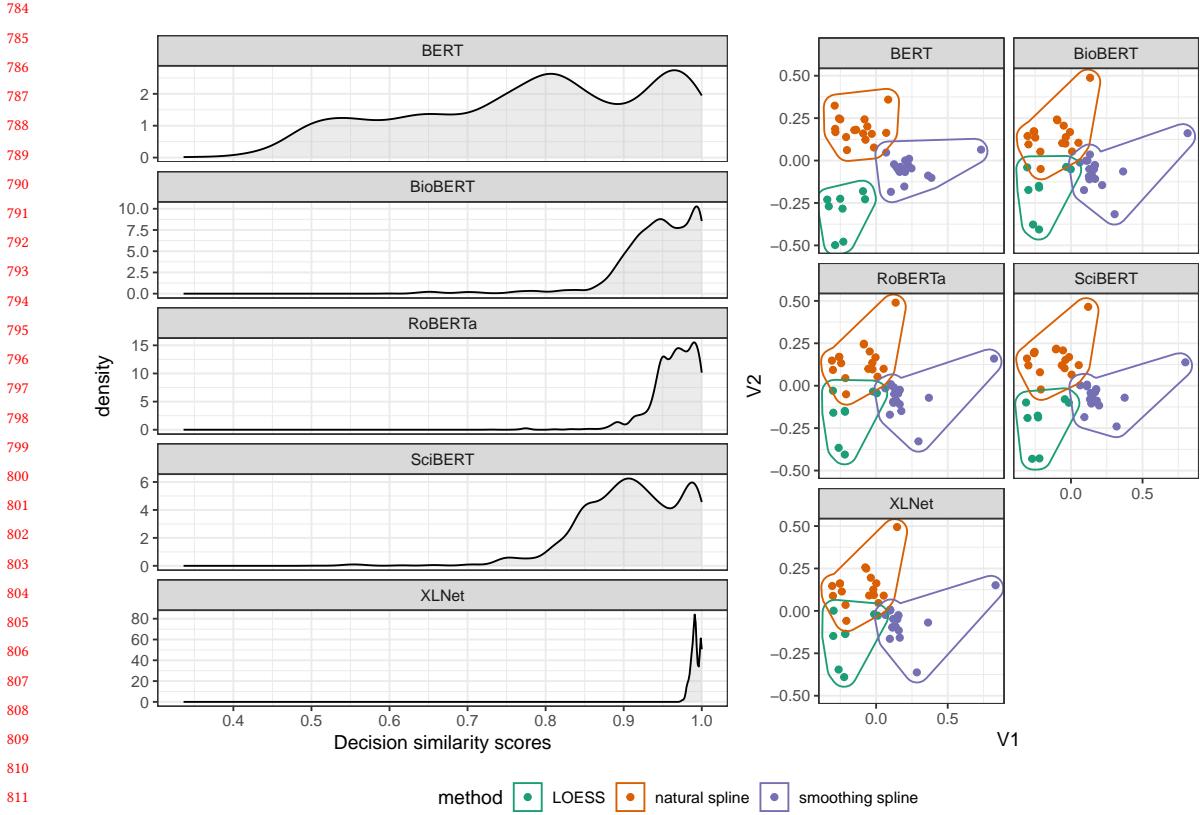


Fig. 5. Distribution of decision similarity (left) and multi-dimensional scaling (MDS) of the paper similarity scores (right) computed for five different text models (BERT, BioBERT, RoBERTa, SciBERT, and XLNet). The default language model, BERT, produces the widest variation across the five models, while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

4.4.3 *Text model.* We have conducted sensitivity analysis on the text model for obtaining the decision similarity score from the Gemini outputs. The tested language models tested include 1) BERT by Google [16], 2) RoBERTa by Facebook AI [35], trained on a larger dataset (160GB v.s. BERT's 15GB), 3) XLNet by Google Brain [56], and two domain-trained BERT models: 4) sciBERT [4], trained on scientific literature, and 5) bioBERT [31], trained on PubMed and PMC data.

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

**833 5 Discussion****834 5.1 Quality of LLM extraction and validation**

835 While the extraction of decisions from literature could be largely automated with LLMs, manual validations remains  
836 essential to ensure the quality of the extracted decisions for further analysis. The quality from the LLM ouput directly  
837 affects the amount of manual effort needed for validation and standardization. Using a default temperature of 1 and  
838 instructing the model to extract original text rather than paraphrase, we find hallucination is not a major issue with  
839 Claude and Gemini for this application. While prompt engineering is used in this work to optimize the prompt for  
840 decision extraction, an alternative is to fine-tune a local model to improve LLM performance. Such approach could  
841 be beneficial for a systematic literature review, although it would require a labelled decision dataset for training and  
842 significantly more training efforts.

**843 5.2 Extracting other types of decisions**

844 As a demonstration, we focus on the modelling decision for the baseline model in the air pollution epidemiology literature.  
845 Analyses in this fields often fit multiple models for different health outcomes. Other models, such as distributed lag  
846 models and multi-pollutant models are also commonly used to estimate relative risks and the interaction among  
847 pollutants. These factors increase the complexity of the decision extraction for LLM, as for additional models, authors  
848 often describe only the differences from the baseline model specification, assuming other decisions remain unchanged.  
849 The LLMs will need to be able to link the decisions across different models and identify the full set of decision for each  
850 model for cross-comparison among papers.

851 Apart from modelling choices, decisions in data pre-processing are also interesting to compare. For example, in  
852 Braga et al. [9], air pollution measures are aggregated from multiple PM10 monitors within the same location into a  
853 single value. Different decisions on how values are extracted, imputed, and aggregated are also shown to affect the  
854 results. However, these decisions are often not well documented in the literature than the modelling decisions, making  
855 it difficult for LLMs to extract them. Proper documentation and reporting of these decisions in future research are  
856 needed before our workflow could be applied to pre-processing decisions.

857 With the advocacy for reproducibility in science, it is expected that more papers will share their code and data. Code  
858 availability can serve as a supplementary source for understanding the choices made in the analysis and cross-check  
859 against the description in the manuscript. However, decision choices could be extracted from the scripts, but the  
860 rationale behind these choices may not be easily discernible given the lack of comments in the current practice.

**861 5.3 Generalizability of the workflow****862 6 Conclusion**

863 In this paper, we aim to study how analysts make decisions in their data analysis practice. While classic interviews  
864 are often conducted in small scale with toy examples, we developed a pipeline for automatically extracting decisions  
865 using LLMs (Claude and Gemini) from scientific literature. We also introduced a method for calculating paper similarity  
866 through comparing the similarities among decisions and the similarity metric can be used as a distance to cluster  
867 papers by their decision choices and visualization with dimension reduction algorithms, such as multidimensional  
868 scaling. We applied this pipeline to a set of air pollution modelling literature that associates daily particulate matter  
869 and daily mortality and hospital admission. From the extracted modelling decisions, we identify the most common  
870 decision choices in this type of analysis and the paper similarity score calculation revealed the three clusters of paper  
871

corresponding to different modelling strategies. These findings are all consistent with the general understanding of the field, as documented in the APHENA project [29] and other methodological comparison studies [42, 50].

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. Futzng and moseyng: 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. Ketzels, 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] 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>.
- [5] 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>.
- [6] Dorothy V. M. Bishop and Charles Hulme. When alternative analyses of the same data come to different conclusions: A tutorial using declaredesign with a worked real-world example. *Advances in Methods and Practices in Psychological Science*, 7(3):25152459241267904, 07 2024. doi: 10.1177/25152459241267904. URL <https://doi.org/10.1177/25152459241267904>. Publisher: SAGE Publications Inc.
- [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. Bottenhorn, Alexander Bowring, Senne Braem, Hayley R. Brooks, Emily G. Brudner, Cristian B. Calderon, Julia A. Camilleri, Jaime J. Castrellon, Luca Cecchetti, Edna C. Cieslik, Zachary J. Cole, Olivier Collignon, Robert W. Cox, William A. Cunningham, Stefan Czoschke, Kamalaker Dadi, Charles P. Davis, Alberto De Luca, Mauricio R. Delgado, Lysia Demetriou, Jeffrey B. Dennison, Xin Di, Erin W. Dickie, Ekaterina Dobryakova, Claire L. Donnat, Juergen Dukart, Niall W. Duncan, Joke Durnez, Amr Eed, Simon B. Eickhoff, Andrew Erhart, Laura Fontanesi, G. Matthew Fricke, Shiguang Fu, Adriana Galván, Remi Gau, Sarah Genon, Tristan Glatard, Enrico Glerean, Jelle J. Goeman, Sergej A. E. Golowin, Carlos González-García, Krzysztof J. Gorgolewski, Cheryl L. Grady, Mikella A. Green, João F. Guassi Moreira, Olivia Guest, Shabnam Hakimi, J. Paul Hamilton, Roeland Hancock, Giacomo Handjaras, Bronson B. Harry, Colin Hawco, Peer Herholz, Gabrielle Herman, Stephan Heunis, Felix Hoffstaedter, Jeremy Hogeveen, Susan Holmes, Chuan-Peng Hu, Scott A. Huettel, Matthew E. Hughes, Vittorio Iacobella, 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 Ertuk Kucukboyaci, Brice A. Kuhl, Sebastian Kupek, Angela R. Laird, Claus Lamm, Robert Langner, Nina Lauharatananahirun, Hongmi Lee, Sangil Lee, Alexander Leemans, Andrea Leo, Elise Lesage, Flora Li, Monica Y. C. Li, Phui Cheng Lim, Evan N. Lintz, Schuyler W. Liphardt, Annabel B. Losecaat Vermeer, Bradley C. Love, Michael L. Mack, Norberto Malpica, Theo Marins, Camille Maumet, Kelsey McDonald, Joseph T. McGuire, Helena Melero, Adriana S. Méndez Leal, Benjamin Meyer, Kristin N. Meyer, Glad Mihai, Georgios D. Mitsis, Jorge Moll, Dylan M. Nielson, Gustav Nilsonne, Michael P. Notter, Emanuele Olivetti, Adrian I. Onicas, Paolo Papale, Kaustubh R. Patil, Jonathan E. Peelle, Alexandre Pérez, Doris Pischedda, Jean-Baptiste Poline, Yanina Prystaika, Shruti Ray, Patricia A. Reuter-Lorenz, Richard C. Reynolds, Emiliano Ricciardi, Jenny R. Rieck, Anais M. Rodriguez-Thompson, Anthony Romyn, Taylor Salo, Gregory R. Samanez-Larkin, Emilio Sanz-Morales, Margaret L. Schlichting, Douglas H. Schultz, Qiang Shen, Margaret A. Sheridan, Jennifer A. Silvers, Kenny Skagerlund, Alec Smith, David V. Smith, Peter Sokol-Hessner, Simon R. Steinkamp, Sarah M. Tashjian, Bertrand Thirion, John N. Thorp, Gustav Tinghög, Loreen Tisdall, Steven H. Tompson, Claudio Toro-Serey, Juan Jesus Torre Tresols, Leonardo Tozzi, Vuong Truong, Luca Turella, Anna E. van ‘t Veer, Tom Verguts, Jean M. Vettel, Sagana Vijayarajah, Khoi Vo, Matthew B. Wall, Wouter D. Weeda, Susanne Weis, David J. White, David Wisniewski, Alba Xifra-Porxas, Emily A. Yearling, Sangsuk Yoon, Rui Yuan, Kenneth S. L. Yuen, Lei Zhang, Xu Zhang, Joshua E. Zosky, Thomas E.

- 937 Nichols, Russell A. Poldrack, and Tom Schonberg. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810):  
938 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.
- 939 [9] Alfésio Luís Ferreira Braga, Antonella Zanobetti, and Joel Schwartz. The lag structure between particulate air pollution and respiratory and  
940 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).
- 941 [10] Richard T. Burnett, Sabit Cakmak, Mark E. Raizenne, David Stieb, Renaud Vincent, Daniel Krewski, Jeffrey R. Brook, Owen Philips, and Haluk  
942 Ozkaynak. The association between ambient carbon monoxide levels and daily mortality in toronto, canada. *Journal of the Air & Waste Management  
943 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>.
- 944 [11] Richard T. Burnett, Stieb ,Dave , Brook ,Jeffrey R. , Cakmak ,Sabit , Dales ,Robert , Raizenne ,Mark , Vincent ,Renaud , , and Tom Dann. Asso-  
945 ciations between short-term changes in nitrogen dioxide and mortality in canadian cities. *Archives of Environmental Health: An International  
946 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:  
947 <https://doi.org/10.3200/AEOH.59.5.228-236> PMID: 16201668.
- 948 [12] Margarita Castillejos, Borja-Aburto ,Victor H. , Dockery ,Douglas W. , Gold ,Diane R. , , and Dana. Loomis. Airborne coarse particles and mortality.  
949 *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>.  
950 Publisher: Taylor & Francis \_eprint: <https://doi.org/10.1080/0895-8378.1987.11463182>.
- 951 [13] Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. Unleashing the potential of prompt engineering for large language models.  
952 *Patterns*, 6(6):101260, 06 2025. doi: 10.1016/j.patter.2025.101260. URL <https://www.sciencedirect.com/science/article/pii/S266389925001084>.
- 953 [14] Chaomei Chen. Citespac ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society  
954 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>.  
955 \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.20317>.
- 956 [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.  
957 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>.
- 958 [16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Naacl-hlt 2019. page 4171–4186, Minneapolis, Minnesota, 06 2019. Association  
959 for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL <https://aclanthology.org/N19-1423/>.
- 960 [17] Douglas W. Dockery, Joel Schwartz, and John D. Spengler. Air pollution and daily mortality: Associations with particulates and acid aerosols.  
961 *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>.
- 962 [18] Marian Dörk, Nathalie Henry Riche, Gonzalo Ramos, and Susan Dumais. Pivotpaths: Strolling through faceted information spaces. *IEEE Transactions  
963 on Visualization and Computer Graphics*, 18(12):2709–2718, 12 2012. doi: 10.1109/TVCG.2012.252. URL <https://ieeexplore.ieee.org/document/6327277>.
- 964 [19] 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  
965 Research Society.
- 966 [20] Elliot Gould, Hannah S. Fraser, Timothy H. Parker, Shinichi Nakagawa, Simon C. Griffith, Peter A. Vesk, Fiona Fidler, Daniel G. Hamilton, Robin N.  
967 Abbey-Lee, Jessica K. Abbott, Luis A. Aguirre, Carles Alcaraz, Irith Aloni, Drew Altschul, Kunal Arekar, Jeff W. Atkins, Joe Atkinson, Christopher M.  
968 Baker, Meghan Barrett, Kristian Bell, Suleiman Kehinde Bello, Iván Beltrán, Bernd J. Berauer, Michael Grant Bertram, Peter D. Billman, Charlie K.  
969 Blake, Shannon Blake, Louis Bliard, Andrea Bonisoli-Alquati, Timothée Bonnet, Camille Nina Marion Bordes, Aneesh P. H. Bose, Thomas Botterill-  
970 James, Melissa Anna Boyd, Sarah A. Boyle, Tom Bradfer-Lawrence, Jennifer Bradham, Jack A. Brand, Martin I. Brengdahl, Martin Bulla, Luc Bussière,  
971 Ettore Camerlenghi, Sara E. Campbell, Leonardo L. F. Campos, Anthony Caravaggi, Pedro Cardoso, Charles J. W. Carroll, Therese A. Catanach,  
972 Xuan Chen, Heung Ying Janet Chik, Emily Sarah Choy, Alec Philip Christie, Angela Chuang, Amanda J. Chunco, Bethany L. Clark, Andrea Contina,  
973 Garth A. Covernton, Murray P. Cox, Kimberly A. Cressman, Marco Crotti, Connor Davidson Crouch, Pietro B. D'Amelio, Alexandra Allison  
974 de Sousa, Timm Fabian Döbert, Ralph Dobler, Adam J. Dobson, Tim S. Doherty, Szymon Marian Drobniak, Alexandra Grace Duffy, Alison B. Duncan,  
975 Robert P. Dunn, Jamie Dunning, Trishna Dutta, Luke Eberhart-Hertel, Jared Alan Elmore, Mahmoud Medhat Elsherif, Holly M. English, David C.  
976 Ensminger, Ulrich Rainer Ernst, Stephen M. Ferguson, Esteban Fernandez-Juricic, Thalita Ferreira-Arruda, John Fieberg, Elizabeth A. Finch, Evan A.  
977 Fiorenza, David N. Fisher, Amélie Fontaine, Wolfgang Forstmeier, Yoan Fourcade, Graham S. Frank, Kathryn A. Freund, Eduardo Fuentes-Lillo,  
978 Sara L. Gandy, Dustin G. Gannon, Ana I. García-Cervigón, Alexis C. Garretson, Xuezhen Ge, William L. Geary, Charly Géron, Marc Gilles, Antje  
979 Girndt, Daniel Glikzman, Harrison B. Goldspiel, Dylan G. E. Gomes, Megan Kate Good, Sarah C. Goslee, J. Stephen Gosnell, Eliza M. Grames, Paolo  
980 Gratton, Nicholas M. Grebe, Skye M. Greener, Maaike Griffioen, Daniel M. Griffith, Frances J. Griffith, Jake J. Grossman, Ali Güncan, Stef Haesen,  
981 James G. Hagan, Heather A. Hager, Jonathan Philo Harris, Natasha Dean Harrison, Sarah Syedia Hasnain, Justin Chase Havird, Andrew J. Heaton,  
982 María Laura Herrera-Chaustré, Tanner J. Howard, Bin-Yan Hsu, Fabiola Iannarilli, Esperanza C. Iranzo, Erik N. K. Iverson, Saheed Olaide Jimoh,  
983 Douglas H. Johnson, Martin Johnsson, Jesse Jorna, Tommaso Jucker, Martin Jung, Ineta Kačergyté, Oliver Kaltz, Alison Ke, Clint D. Kelly, Katharine  
984 Keegan, Friedrich Wolfgang Keppeler, Alexander K. Killion, Dongmin Kim, David P. Kochan, Peter Korsten, Shan Kothari, Jonas Kuppler, Jillian M.  
985 Kusch, Małgorzata Lagisz, Kristen Marianne Lalla, Daniel J. Larkin, Courtney L. Larson, Katherine S. Lauck, M. Elise Lauterbur, Alan Law, Don-Jean  
986 Léandri-Breton, Jonas J. Lembrechts, Kiara L'Herpiniere, Eva J. P. Lievens, Daniela Oliveira de Lima, Shane Lindsay, Martin Luquet, Ross MacLeod,  
987 Kirsty H. Macphie, Kit Magellan, Magdalena M. Mair, Lisa E. Malm, Stefano Mammola, Caitlin P. Mandeville, Michael Manhart, Laura Milena  
988

- 989 Manrique-Garzon, Elina Mäntylä, Philippe Marchand, Benjamin Michael Marshall, Charles A. Martin, Dominic Andreas Martin, Jake Mitchell  
 990 Martin, April Robin Martinig, Erin S. McCallum, Mark McCauley, Sabrina M. McNew, Scott J. Meiners, Thomas Merkling, Marcus Michelangeli,  
 991 Maria Moiron, Bruno Moreira, Jennifer Mortensen, Benjamin Mos, Taofeek Olatunbosun Muraina, Penelope Wrenn Murphy, Luca Nelli, Petri  
 992 Niemelä, Josh Nightingale, Gustav Nilsonne, Sergio Nolazco, Sabine S. Nooten, Jessie Lanterman Novotny, Agnes Birgitta Olin, Chris L. Organ,  
 993 Kate L. Ostevik, Facundo Xavier Palacio, Matthieu Paquet, Darren James Parker, David J. Pascall, Valerie J. Pasquarella, John Harold Paterson, Ana  
 994 Payo-Payo, Karen Marie Pedersen, Grégoire Perez, Kayla I. Perry, Patrice Pottier, Michael J. Proulx, Raphaël Proulx, Jessica L. Pruitt, Veronarindra  
 995 Ramananjato, Finaritra Tolotra Randimbiarison, Onja H. Razafindratsima, Diana J. Rennison, Federico Riva, Sepand Riyahi, Michael James Roast,  
 996 Felipe Pereira Rocha, Dominique G. Roche, Cristian Román-Palacios, Michael S. Rosenberg, Jessica Ross, Freya E. Rowland, Deusdedith Rugemalila,  
 997 Avery L. Russell, Suvi Ruuskanen, Patrick Saccone, Asaf Sadeh, Stephen M. Salazar, Kris Sales, Pablo Salmón, Alfredo Sánchez-Tójár, Leticia Pereira  
 998 Santos, Francesca Santostefano, Hayden T. Schilling, Marcus Schmidt, Tim Schmoll, Adam C. Schneider, Allie E. Schroeder, Nicolas  
 999 Schtickzelle, Nick L. Schultz, Drew A. Scott, Michael Peter Scroggie, Julie Teresa Shapiro, Nitika Sharma, Caroline L. Shearer, Diego Simón, Michael I.  
 1000 Sitvarin, Fabricio Luiz Skupien, Heather Lea Slimm, Grania Polly Smith, Jeremy A. Smith, Rahel Sollmann, Kaitlin Stack Whitney, Shannon Michael  
 1001 Still, Erica F. Stuber, Guy F. Sutton, Ben Swallow, Conor Claverie Taff, Elina Takola, Andrew J. Tanentzap, Rocío Tarjuelo, Richard J. Telford,  
 1002 Christopher J. Thawley, Hugo Thierry, Jacqueline Thomson, Svenja Tidau, Emily M. Tompkins, Claire Marie Tortorelli, Andrew Trlica, Biz R.  
 1003 Turnell, Lara Urban, Stijn Van de Vondel, Jessica Eva Megan van der Wal, Jens Van Eeckhoven, Francis van Oordt, K. Michelle Vanderwel, Mark C.  
 1004 Vanderwel, Karen J. Vanderwolf, Juliana Vélez, Diana Carolina Vergara-Florez, Brian C. Verrelli, Marcus Vinícius Vieira, Nora Villamil, Valerio  
 1005 Vitali, Julien Vollering, Jeffrey Walker, Xanthe J. Walker, Jonathan A. Walter, Paweł Waryszak, Ryan J. Weaver, Ronja E. M. Wedegärtner, Daniel L.  
 1006 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>.
- [21] Florian Heimerl, Qi Han, Steffen Koch, and Thomas Ertl. Citerivers: Visual analytics of citation patterns. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):190–199, 01 2016. doi: 10.1109/TVCG.2015.2467621. URL <https://ieeexplore.ieee.org/document/7192685/authors>.
- [22] Wei Huang, Jianguo Tan, Haodong Kan, Ni Zhao, Weimin Song, Guixiang Song, Guohai Chen, Lili Jiang, Cheng Jiang, Renjie Chen, and Bingheng 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.scitotenv.2009.02.019. URL <https://linkinghub.elsevier.com/retrieve/pii/S004896970900165X>.
- [23] Nick Huntington-Klein, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pralhad Burli, Naibin Chen, Paul Grieco, Godwin Ekpe, Todd Pugatch, Martin Saavedra, and Yaniv Stopnitzky. The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry*, 59(3):944–960, 2021. doi: 10.1111/ecin.12992. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/ecin.12992>. \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecin.12992>.
- [24] Petra Isenberg, Tobias Isenberg, Michael Sedlmair, Jian Chen, and Torsten Möller. Visualization as seen through its research paper keywords. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):771–780, 01 2017. doi: 10.1109/TVCG.2016.2598827. URL <https://ieeexplore.ieee.org/document/7539364>.
- [25] 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>.
- [26] Haodong 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>.
- [27] 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.
- [28] 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).
- [29] 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.
- [30] 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/>.
- [31] Jinyuk 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>.
- [32] 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/>
- 1040 Manuscript submitted to ACM

- 1041                   8805460/.
- 1042 [33] Yang Liu, Tim Althoff, and Jeffrey Heer. Paths explored, paths omitted, paths obscured: Decision points & selective reporting in end-to-end  
1043 data analysis. CHI '20, page 1–14, New York, NY, USA, 04 2020. Association for Computing Machinery. doi: 10.1145/3313831.3376533. URL  
1044 <https://dl.acm.org/doi/10.1145/3313831.3376533>.
- 1045 [34] Yang Liu, Alex Kale, Tim Althoff, and Jeffrey Heer. Boba: Authoring and visualizing multiverse analyses. *IEEE Transactions on Visualization and*  
1046 *Computer Graphics*, 27(2):1753–1763, 02 2021. doi: 10.1109/TVCG.2020.3028985. URL <https://ieeexplore.ieee.org/document/9216579/>.
- 1047 [35] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov.  
1048 Roberta: A robustly optimized bert pretraining approach. doi: 10.48550/arXiv.1907.11692.
- 1049 [36] Elena López-Villarrubia, Ferran Ballester, Carmen Iniguez, and Nieves Peral. Air pollution and mortality in the canary islands: a time-series analysis.  
1050 *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  
1051 PMCID: PMC2843667.
- 1052 [37] T F Mar, G A Norris, J Q Koenig, and T V Larson. Associations between air pollution and mortality in phoenix, 1995–1997. *Environmental*  
1053 *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:  
1054 Environmental Health Perspectives.
- 1055 [38] Suresh H. Moolgavkar. Air pollution and hospital admissions for diseases of the circulatory system in three u.s. metropolitan areas. *Journal of the Air &*  
1056 *Waste Management Association*, 50(7):1199–1206, 07 2000. doi: 10.1080/10473289.2000.10464162. URL <https://doi.org/10.1080/10473289.2000.10464162>.  
1057 Publisher: Taylor & Francis.
- 1058 [39] Suresh H. Moolgavkar. Air pollution and daily mortality in two u.s. counties: Season-specific analyses and exposure-response relationships.  
1059 *Inhalation Toxicology*, 15(9):877–907, 01 2003. doi: 10.1080/08958370390215767. URL <https://doi.org/10.1080/08958370390215767>. Publisher: Taylor &  
1060 Francis.
- 1061 [40] Arpit Narechania, Alireza Karduni, Ryan Wesslen, and Emily Wall. Vitality: Promoting serendipitous discovery of academic literature with  
1062 transformers & visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):486–496, 01 2022. doi: 10.1109/TVCG.2021.3114820.  
1063 URL <https://ieeexplore.ieee.org/document/9552447/>.
- 1064 [41] Bart Ostro, Rachel Broadwin, Shelley Green, Wen-Ying Feng, and Michael Lipsett. Fine particulate air pollution and mortality in nine california  
1065 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.
- 1066 [42] Roger D. Peng, Francesca Dominici, and Thomas A. Louis. Model choice in time series studies of air pollution and mortality. *Journal of the Royal*  
1067 *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>.
- 1068 [43] 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,  
1069 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>.
- 1070 [44] Abhraneel Sarma, Alex Kale, Michael Moon, Nathan Taback, Fanny Chevalier, Jessica Hullman, and Matthew Kay. multiverse: Multiplexing  
1071 alternative data analyses in r notebooks (version 0.6.2). *OSF Preprints*, 2021. URL <https://github.com/MUCollective/multiverse>.
- 1072 [45] Marko Sarstedt, Susanne J. Adler, Christian M. Ringle, Gyeongcheol Cho, Adamantios Diamantopoulos, Heungsun Hwang, and Benjamin D.  
1073 Liengaard. Same model, same data, but different outcomes: Evaluating the impact of method choices in structural equation modeling. *Journal of*  
1074 *Product Innovation Management*, 41(6):1100–1117, 2024. doi: 10.1111/jpim.12738. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/jpim.12738>.  
1075 \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/jpim.12738>.
- 1076 [46] 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.
- 1077 [47] R. Silberzahn, E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtrey, Š. Bahník, F. Bai, C. Bannard, E. Bonnier, R. Carlsson, F. Cheung,  
1078 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,  
1079 T. J. Heaton, K. Hederos, M. Heene, A. J. Hofelich Mohr, F. Högden, K. Hui, M. Johannesson, J. Kalodimos, E. Kaszubowski, D. M. Kennedy, R. Lei,  
1080 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,  
1081 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,  
1082 L. Stefanutti, S. Tauber, J. Ullrich, M. Vianello, E.-J. Wagenamakers, M. Witkowiak, S. Yoon, and B. A. Nosek. Many analysts, one data set: Making  
1083 transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, 1(3):337–356, 09 2018. doi:  
1084 10.1177/2515245917747646. URL <https://doi.org/10.1177/2515245917747646>. Publisher: SAGE Publications Inc.
- 1085 [48] Jan Simson, Fiona Draxler, Samuel Mehr, and Christoph Kern. Preventing harmful data practices by using participatory input to navigate the  
1086 machine learning multiverse. CHI '25, page 1–30, New York, NY, USA, 04 2025. Association for Computing Machinery. doi: 10.1145/3706598.3713482.  
1087 URL <https://dl.acm.org/doi/10.1145/3706598.3713482>.
- 1088 [49] Imad Tbahrith, Christine Chichester, Frédérique Lisacek, and Patrick Ruch. Using argumentation to retrieve articles with similar citations: An  
1089 inquiry into improving related articles search in the medline digital library. *International Journal of Medical Informatics*, 75(6):488–495, 06 2006. doi:  
1090 10.1016/j.ijmedinf.2005.06.007. URL <https://www.sciencedirect.com/science/article/pii/S1386505605000894>.
- 1091 [50] G. Touloumi, E. Samoli, M. Pipiou, A. Le Tertre, R. Atkinson, and K. Katsouyanni. Seasonal confounding in air pollution and health time-series  
1092 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>.

- 1093 com/doi/abs/10.1002/sim.2681. \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.2681>.
- 1094 [51] Kayo Ueda, Nitta ,Hiroshi , Ono ,Masaji , , and Ayano Takeuchi. Estimating mortality effects of fine particulate matter in japan: A comparison of time-  
1095 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.  
1096 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>.
- 1097 [52] 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>.
- 1098 [53] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes,  
1099 Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson,  
1100 Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal  
of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.
- 1101 [54] Hadley Wickham, Joe Cheng, and Aaron Jacobs. *ellmer: Chat with Large Language Models*, 2025. URL <https://CRAN.R-project.org/package=ellmer>.  
1102 R package version 0.1.1.
- 1103 [55] Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. Large language  
1104 models for generative information extraction: A survey. doi: 10.48550/arXiv.2312.17617.
- 1105 [56] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for  
1106 language understanding. doi: 10.48550/arXiv.1906.08237.
- 1107  
1108  
1109  
1110  
1111  
1112  
1113  
1114  
1115  
1116  
1117  
1118  
1119  
1120  
1121  
1122  
1123  
1124  
1125  
1126  
1127  
1128  
1129  
1130  
1131  
1132  
1133  
1134  
1135  
1136  
1137  
1138  
1139  
1140  
1141  
1142  
1143  
1144 Manuscript submitted to ACM