

1 **The Name of the Title Is Hope**

2
3 ANONYMOUS AUTHOR(S)

4
5 bla blabla

6
7 CCS Concepts: • **Applied computing** → *Document analysis; Environmental sciences;* • **Human-centered computing** → **HCI**
8 **theory, concepts and models.**

9 Additional Key Words and Phrases: Large language models

10 **ACM Reference Format:**

11 Anonymous Author(s). 2025. The Name of the Title Is Hope. In *Proceedings of CHI Conference on Human Factors in Computing Systems (CHI'26)*. ACM, New York, NY, USA, 18 pages. <https://doi.org/XXXXXX.XXXXXXX>

- 12
13
14
15 • Something about “analysis review” - Roger thinks it’s a better to have a new word for this.
16 • provide a baseline understand - place to start
17 • demonstrate - analytically homogeneous - the table won’t look like that

18
19 **1 Introduction**

20
21 Decisions are everywhere in data analysis, from the initial data collection, data pre-processing to the modelling
22 choices. These decisions will impact the final output of the data analysis, which may lead to different conclusions
23 and policy recommendations. When such flexibility can be misused—through practices such as p-hacking, selective
24 reporting, or unjustified analytical adjustments—it can inflate effect sizes or produce misleading results that meet
25 conventional thresholds for statistical significance. They have been demonstrated through many-analysts experiments,
26 where independent teams analyzing the same dataset to answer a pre-defined research question often arrive at markedly
27 different conclusions. These practices not only compromise the validity of individual studies but also threaten the
28 broader credibility of statistical analysis and scientific research as a whole.

29
30 Multiple recommendations have been proposed to improve data analysis practices, such as pre-registration and
31 multiverse analysis. Bayesian methods also offer a different paradigm to p-value driven inference for interpreting
32 statistical evidence. Most empirical studies of data analysis practices focus on specially designed and simplified analysis
33 scenarios. While informative, these setups may not adequately capture the complexity of the data analysis with
34 significant policy implications. [In practice, studying the data analysis decisions with actual applications is challenging.]
35 Analysts may no longer be available for interviews due to job changes, and even when they are, recalling the full set
36 of decisions and thinking process made during the analysis is often infeasible. Moreover, only until the last decades,
37 analysis scripts and reproducible materials were not commonly required by journals for publishing. [As a result, it
38 remains challenging to study how analytical decisions are made.]

39
40 In this work, we focus on a specific class of air pollution modelling studies that estimate the effect size of particulate
41 matter (PM2.5 or PM10) on mortality, typically using Poisson regression or generalized additive models (GAMs).

42 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
43 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
44 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
45 redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

46 © 2025 ACM.

47 Manuscript submitted to ACM

48
49 Manuscript submitted to ACM

53 While individual modelling choices vary, these studies often share a common structure: they adjust for meteorological
 54 covariates such as temperature and humidity, apply temporal or spatial treatments, like including lagged variables and
 55 may estimate the effect by city or region before combining results. Because these studies investigate similar scientific
 56 questions using a shared modelling framework, they form a natural many-analyst setting. This allows us to examine, in
 57 a real-world context, the range of analytical decisions made by different researchers addressing the same underlying
 58 question.
 59

60 In this work, we develop a structured tabular format to record the analytical decisions made by researchers in the air
 61 pollution modelling literature. Using large language models (LLMs), we automate the extraction of these decisions from
 62 published papers. This allows us to treat decisions as data – allowing us to track them over time, compare methodology
 63 across papers, and query commonly used approaches. We further introduce a workflow to cluster studies based on
 64 decision similarity, revealing three distinct groups of papers that reflect the modelling strategies differ in the European
 65 and U.S. studies, which offers a new way to visualize the field in the air pollution mortality modelling.
 66

67 The rest of the paper is organized as follows. In Section 2, we review the background on data analysis decisions.
 68 Section 3 describes the data structure for recording decisions, the use of large language models to process research
 69 papers, and the validation of LLM outputs. In Section 4, we present the method for calculating paper similarity based
 70 on decision similarities. Section 5 reports the finding of our analysis, including the clustering of papers according to
 71 similarity scores and sensitivity analyses related to LLM providers, prompt engineering, and LLM parameters. Finally,
 72 Section 6 discusses the implications of our study.
 73

74 2 Background

75 2.1 Decisions in data analysis

76 **Question** Is “decision” going to be confusing with “decision-making” in decision theory

77 A data analysis is a process of making choices at each step, from the initial data collection to model specification, and
 78 post-processing. Each decision represents a branching point where analysts choose a specific path to follow, and the
 79 vast number of possible choices analysts can take forms what Gelman and Loken [7] describe as the “garden of forking
 80 paths”. While researchers may hope their inferential results are robust to the specific path taken through the garden,
 81 in practice, different choices can lead to substantially different conclusions. This has been empirically demonstrated
 82 through “many analyst experiments”, where independent research groups analyze the same dataset to the same answer
 83 using their chosen analytic approach. A classic example is Silberzahn et al. [22], where researchers reported an odds
 84 ratio from 0.89 to 2.93 for the effect of soccer players’ skin tone on the number of red cards awarded by referees. Similar
 85 variability has been observed in structural equation modeling [21], applied microeconomics [10], neuroimaging [5], and
 86 ecology and evolutionary biology [8]. Many studies have been conducted on a relatively smaller scale to interviews of
 87 analysts and researchers on data analysis practices [1, 12, 16], visualization of the decision process through the analytic
 88 decision graphics (ADG) [17]. Recently, Simson et al. [23] describes a participatory approach to decisions choices in
 89 fairness ML algorithms.
 90

91 Software tools have also developed to incorporate potential alternatives in the analysis workflow, including the
 92 DeclareDesign package [4] and the multiverse package [20]. The DeclareDesign package [4] introduces the MIDA
 93 framework for researchers to declare, diagnose, and redesign their analyses to produce a distribution of the statistic of
 94 interest, which has been applied in the randomized controlled trial study [3]. The multiverse package [20] provides
 95

105 a framework for researchers to systematically explore how different choices affect results and to report the range of
 106 plausible outcomes that arise from alternative analytic paths.
 107

108 **TODO** Something about the context on air pollution mortality modelling @ Roger

109 **3 Extracting decisions from data analysis**

110 **3.1 Decisions in data analysis**

111 Decisions occur throughout the entire data analysis process – from the selection of variables and data source, to
 112 pre-processing steps to prepare the data for modelling, to the model specification and variable inclusion. In this work,
 113 we focus specifically on modelling decisions in the air pollution mortality modelling literature. These include the
 114 choice of modelling approach, covariate inclusion and smoothing, and specifications of spatial and temporal structure.
 115 Consider the following excerpt from Ostro et al. [19]:

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

119 This sentence encode the following components of a decision:

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

126 The decision above is regarding a certain parameter in the statistical method, we categorize this as a “parameter”
 127 type decisions. Other types of decisions - such as spatial modelling structure or the inclusion of temporal lags - may
 128 not include an explicit method or parameter, but still reference a variable and rationale, which we will provide further
 129 examples below.

130 To record these decisions, we follow the tidy data principle [24], where each variable should be in a column, each
 131 observation in a row. In our context, each row represents a decision made by the authors of a paper and an analysis
 132 often include multiple decisions. To retain the original context of the decision, we extract the original text in the paper,
 133 without paraphrase or summarization, from the paper. Below we present an example of how to structure the decisions
 134 made in a paper, using the paper by Ostro et al. [19]:

Paper	ID	Model	variable	method	parameter	type	reason	decision
ostro	1	Poisson regression	temperature	smoothing spline	degree of freedom	parameter	NA	3 degree of freedom
ostro	2	Poisson regression	temperature	smoothing spline	degree of freedom	temporal	NA	1-day lag

Paper	ID	Model	variable	method	parameter	type	reason	decision
ostro	3	Poisson regression	relative humidity	LOESS	smoothing parameter	parameter	to minimize Akaike's Information Criterion	NA
ostro	4	Poisson regression	model	NA	NA	spatial	to account for variation among cities	separate regression models fit in each city

Most decisions in the published papers are not explicitly stated, this could due to the coherence and conciseness of the writing or authors' decision to include only necessary details. Here, we identify a few common anomalies where decisions may be combined or omit certain fields:

1. **Authors may combine multiple decisions into a single sentence** for coherence and conciseness of the writing. Consider the following excerpt from Ostro et al. [19]:

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

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

2. **The justification does not directly address the decision choice.** In the example above, the stated rationale ("and are likely to vary over time in concert with air pollution levels") supports the general inclusion of temporal lags but does not justify the specific choice of 1-day lag over alternatives, such as 2-day average of lags 0 and 1 (lag01) and single-day lag of 2 days (lag2). As such, the reason field should be recorded as NA.

3. **Some decisions may be omitted because they are data-driven.** For instance, Katsouyanni et al. [14] states: 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.

In this case, while the method of selection (minimizing AIC) is specified, the actual degree of freedom used is not. Such data-driven decisions may be recorded with "NA" in the decision field, but the reason field should still be recorded as "by minimizing Akaike's information criterion"

4. **Information required to interpret the decision may be distributed across multiple sections.** In the previous example, "weather variables" refers to mean temperature and relative humidity, as defined earlier in the text. This requires cross-referencing across sections to identify the correct variables associated with each modeling choice.

209 3.2 Automatic reading of literature with LLMs

210 **TODO:** Prompt engineering: these models may paraphrase or hallucinate unless explicitly told not to since it is
211 generative in nature based on the predicted probability of the next word from the text and the instruction

212 While decisions can be extracted manually from the literature, this process is labor-intensive and time-consuming.
213 Recent advances in Large Language Models (LLMs) have demonstrated potential for automating the extraction of
214 structured information from unstructured text [ref]. In this work, we use LLMs to automatically identify decisions
215 made by authors during their data analysis processes.

216 Text recognition from PDF document relies on Optical Character Recognition (OCR) to convert scanned images into
217 machine-readable text – capability currently offered by Anthropic Claude and Google Gemini. We instruct the LLM
218 to generate a markdown file containing a JSON block that records extracted decisions, which can then be read into
219 statistical software for further analysis. The exact prompt feed to the LLM is provided in the Appendix. The `ellmer`
220 package [25] in R is used to connect to the Gemini and Claude API, providing the PDF attachment and the prompt in a
221 markdown file as inputs.

222 3.3 Review the LLM output

- 223 • **TODO** something about result validation of LLM output

224 The shiny app is designed to provide users a visual interface to review and edit the decisions extracted by the LLM
225 from the literature. The app allows three actions from the users: 1) *overwrite* – modify the content of a particular
226 cell, equivalently `dplyr::mutate(xxx = ifelse(CONDITION, "yyy" , xxx))`, 2) *delete* – remove a particular cell,
227 `dplyr::filter(!(CONDITION))`, and 3) *add* – manually enter a decision, `dplyr::bind_rows()`. Figure 1 illustrates
228 the *overwrite* action in the Shiny application, where users interactively filter the data and preview the rows affected by
229 their edits—in this case, changing the model entry from “generalized additive Poisson time series regression” to the
230 less verbose “Poisson regression”. Upon confirmation, the corresponding tidyverse code is generated, and users can
231 download the edited table and incorporate the code into their R script.

261 Edit decision table output
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312

1

2

3

4

Fig. 1. The Shiny application interface for editting Large Language Model (LLM)-generated decisions (overwrite, delete, and add). (1) the default interface after loading the input CSV file. (2) The table view will update interactively upon the user-defined filter condition – expressed using `dplyr::filter()` syntax (e.g., `paper == anderson2008size`), (3) The user edits the model column to “Poisson regression” and applies the change by clicking the Apply changes button. The table view updates to reflect the changes (4) After clicking the Confirm button, the corresponding `tidyverse` code is generated, and the table view returns to its original unfiltered view. The edited data can be downloaded by clicking the Download CSV button.

4 Calculating paper similarity

Once the decisions have been extracted and validated, this opens up a structured data for analyzing these information. For example, we can compare whether author’s choices at different times changes, or across decisions varies at different regions. In this section, we present a method to calculate paper similarity based on the decisions shared in the paper pairs. The goal is to construct a distance metric based on similarity of the decision choice among papers that could be further used for clustering paper based on choices made by different authors in the literature. An overview of the method is illustrated in Figure 2.

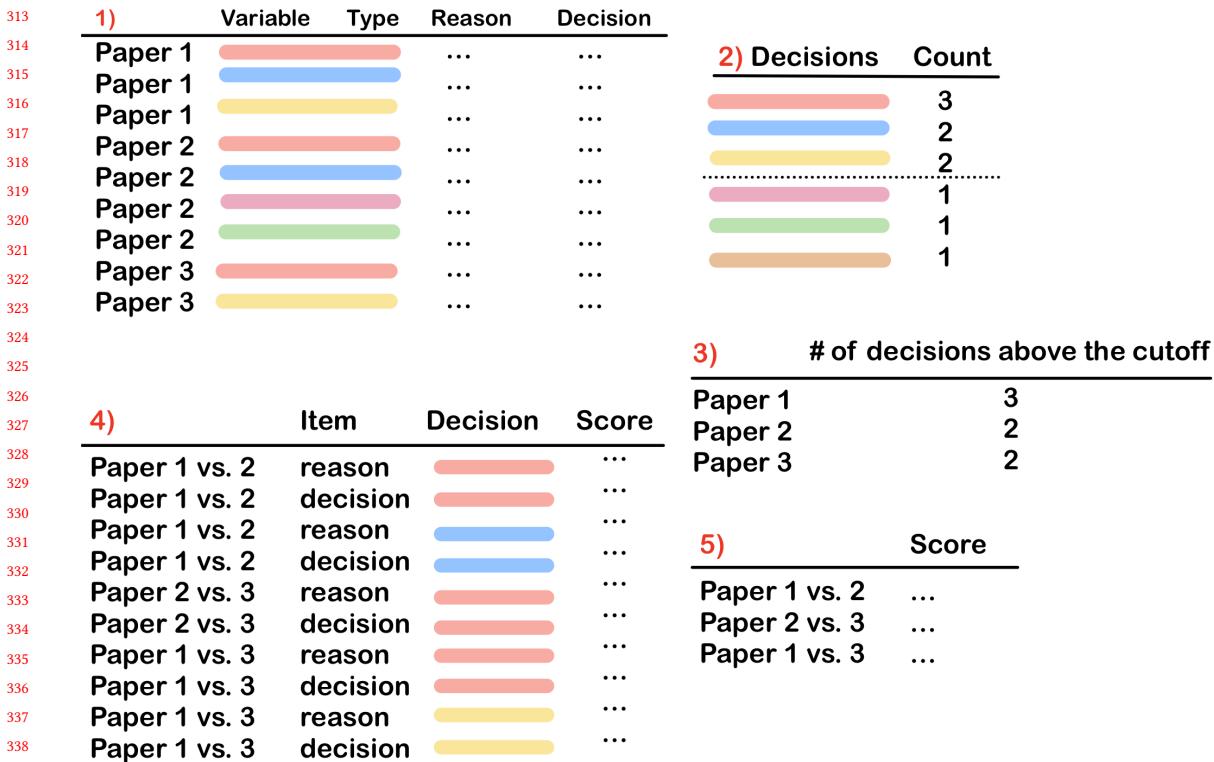


Fig. 2. Workflow for calculating paper similarity based on decision choices: (1) standardize variable names, (2) identify most frequent variable-type decisions across all papers, (3) identify papers with at least x identified decisions, (4) calculate decisions similarity score on the *decision* and *reason* fields using transformer language models, e.g. BERT, (5) calculate paper similarity score based on aggregating decision similarity scores.

- TODO some discussion on what it means by for two papers to be similar based on decisions.

The calculation of paper similarity is based on the similarity of decisions shared by each paper pair. A decision comparable in two papers are the ones that share the same variable and type, e.g. temperature and parameter (a decisions on the choosing the statistical method *parameter* for the *temperature* variable), or humidity and temporal (any *temporal* treatment, e.g. choice of lag value for the *humidity* variable). While many decisions share a similar variable, different authors may refer to them with slightly different names, such as “mean temperature” and “average temperature”, hence variable names are first standardized to a common set of variable names. For example, “mean temperature” and “average temperature” are both standardized to “temperature”. Notice that “dewpoint temperature” is standardized into “humidity” since it is a proxy of temperature to achieve a relative humidity (RH) of 100%. For literature with a common theme, there is usually a set of variables that shared by most papers and additional variables are justified in individual research.

For our air pollution mortality modelling literature, we standardize the following variable names:

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

- 365 • **time**: “date”, “time”, “trends”, “trend”

366 Depending on the specific pairs, papers have varied number of decisions that can be compared and aggregated. While
 367 paper similarities can be computed for all paper pairs, using the similarity of one or two pair of decisions to represent
 368 paper similarity is less ideal. Hence, before calculating the text similarity of decisions, we also include two optional
 369 steps to identify and subset the most frequent decisions across papers, and to retain only papers that report more than
 370 a certain number of frequent decisions. Research questions in different fields may have different levels of homogeneity,
 371 depending on the maturity of the field and for air pollution mortality modelling, it is helpful to focus on decisions and
 372 papers that share a substantial number of decisions.

373 To assign numerical value for the similarity of reason, we use a transformer language model, such as BERT, to
 374 measure the semantic text similarity between the decision itself and its justification. The decision similarity is calculated
 375 by comparing the *decision* and *reason* fields of the decisions in each paper pair. To obtain paper similarity, we average
 376 the decision similarities across all decisions in each paper pair and other method can be customized for aggregation.
 377 The resulting paper similarity score can be used as a distance matrix to cluster papers based on their decision choices to
 378 understand the common practices in the investigated literature.

395 5 Results

396 5.1 Air pollution mortality modelling

397 The results follows examines [x] papers for modelling the effect of particulate matters on mortality based on Gemini
 398 for parsing the decision choices. The results from Anthropic Claude is reported in Section 5.2.

400 Specify how much of validation and review has been done

401 Decision quality summary

- 403 • missingness of the reason and decisions for the paper - how often papers report decisions
- 404 • look at for one type of decision (time) - what are the choices made by different papers
- 405 • look at whether decisions changes across time (cluster diagram with year)
- 406 • Visualize the decision database: apply clustering algorithm and visualize the clusters
- 407 • a characterization of the field, what are the common variables included, what smoothing methods are used,
 408 what are the options for temporal lags often considered, how are models generally estimated spatially.

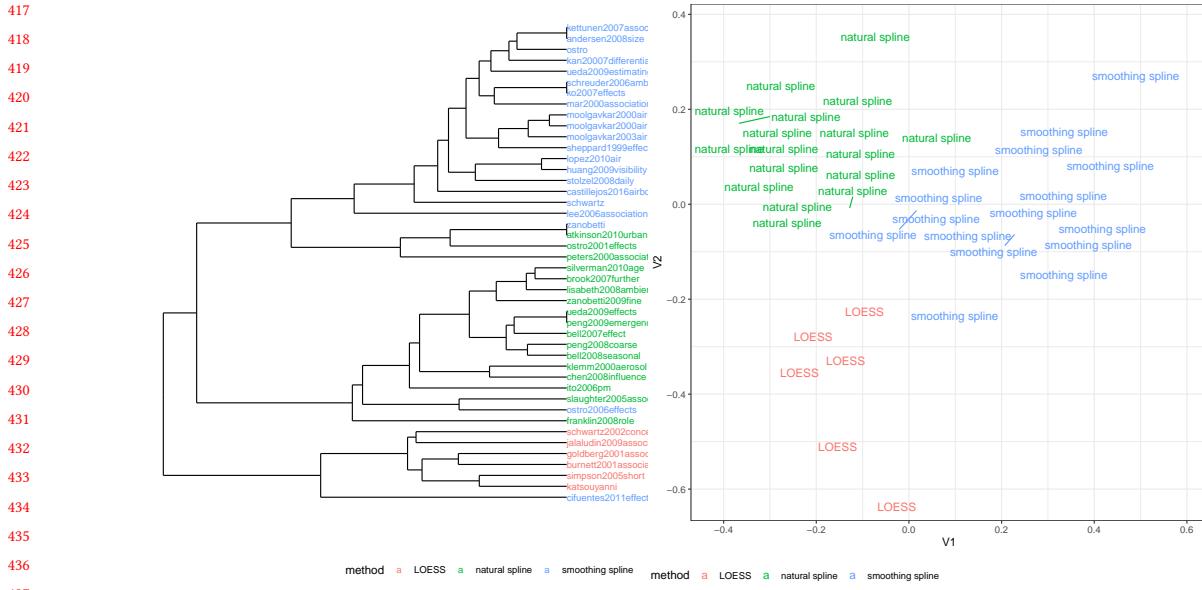


Fig. 3. bla bla bla

5.2 Sensitivity analysis

sensitivity of the pipeline: 1) LLM, 2) text model, 3) LLM parameters

- A section on reproducibility of LLM outputs: prompt experiment (see if there are papers discussing this: <https://arxiv.org/pdf/2406.06608>)

5.2.1 LLM provider. Input of PDF, require OCR, which is only supported by Antropic Claude and Google Gemini. Comparison of the specification from the their sites Claude is decoder only, Gemini is an encoder-decoder model

Identify the proportion of the common decision extracted by Claude and Gemini

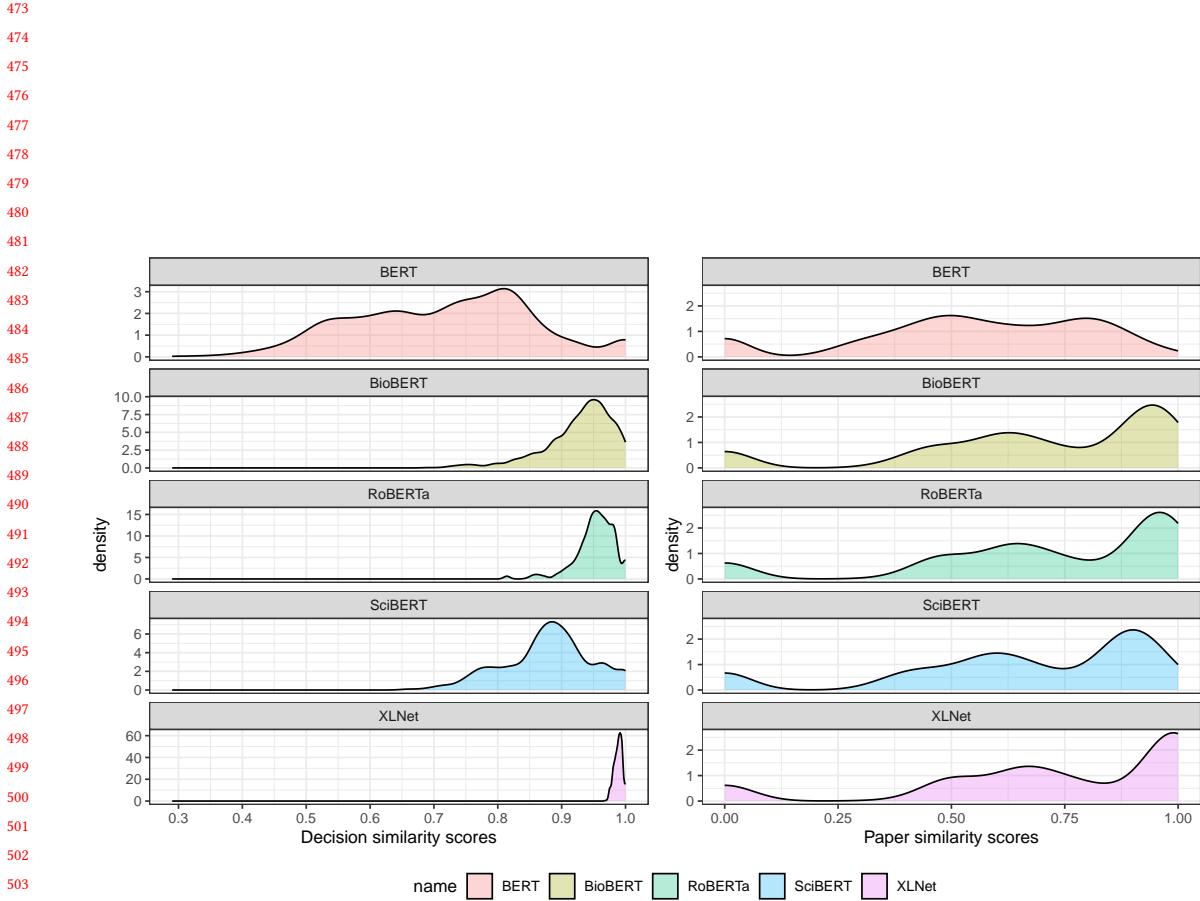
Compare the clustering by Claude and Gemini

- Claude tends to give more robust output while Gemini has a higher accuracy.
- Given examples of the failure of LLM models for parsing and examples where authors are unclear about the delivery

5.2.2 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) standard BERT (google-bert/bert-base-uncased) [6], 2) Roberta (FacebookAI/roberta-base) [18], which is trained on a much larger dataset (160GB v.s. BERT's 15GB), 3) xlnet by Google Brain (xlnet/xlnet-base-cased) [26], and two domain-trained BERT models: 4) sciBert (allenai/scibert_scivocab_uncased) [2], trained on scientific literature, and 5) bioBert (dmis-lab/biobert-large-cased-v1.1-squad) [15], trained on PubMed and PMC data.

Figure 4 presents the distribution of the decision similarity (left) and paper similarity (right) 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. These scores are not comparable across models since the difference of the underlying

469 transformer architecture. However, the paper similarity scores from each model are comparable and Figure 5 shows the
 470 multi-dimensional scaling (MDS) of the paper similarity scores from each text model: all showing a similar clustering
 471 pattern of the three main smoothing methods.
 472



506 Fig. 4. Distribution of decision similarity (left) and paper similarity (right) scores for five different text models (BERT, BioBERT,
 507 RoBERTa, SciBERT, and XLNet). The default language model, BERT, produces the widest variation across the five models, while
 508 the similarity scores form XLNet are all close to 1. The model BioBERT, RoBERTa, and SciBERT yield decision scores mostly
 509 between 0.7 to 1.

510
 511
 512
 513
 514
 515
 516
 517
 518
 519

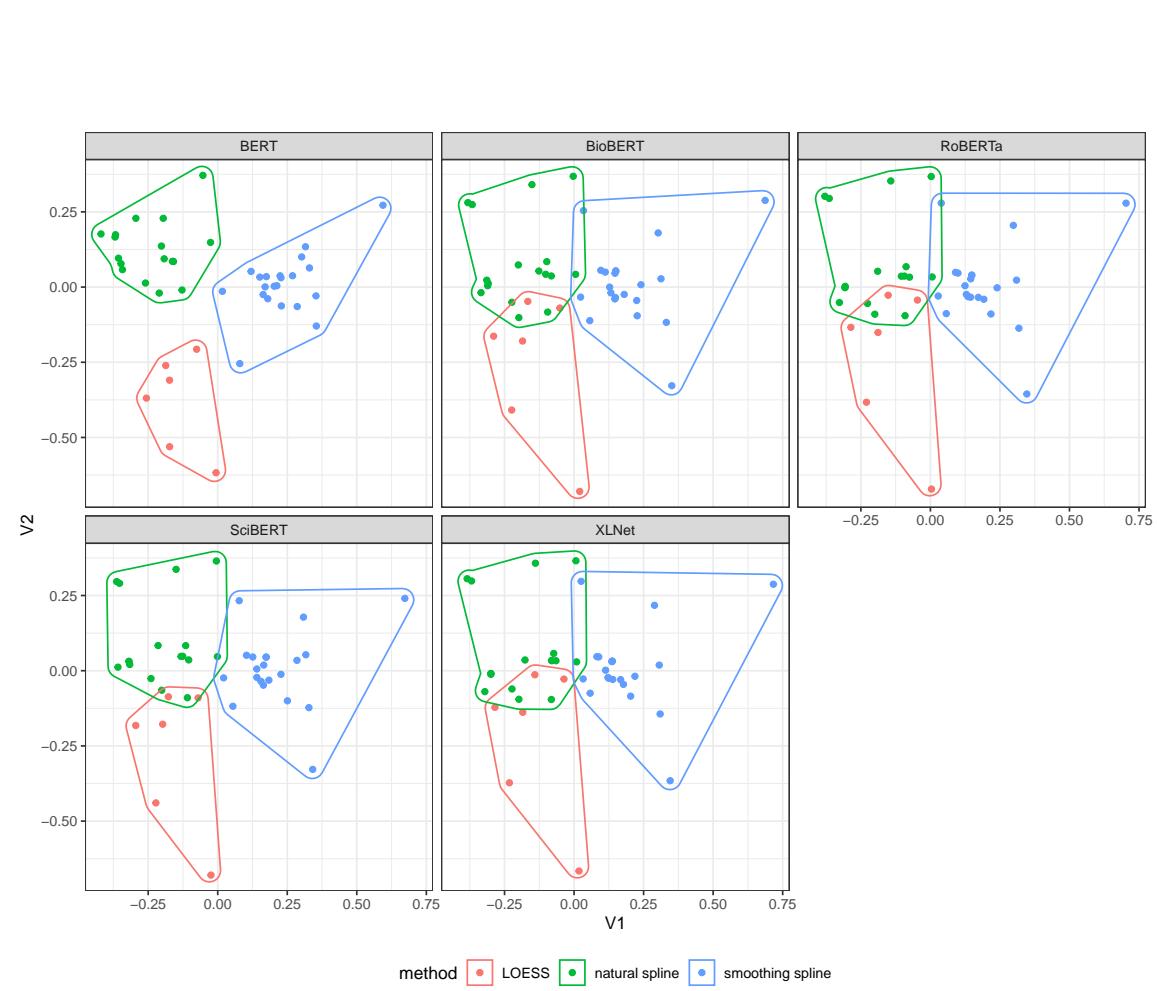


Fig. 5. The multi-dimensional scaling (MDS) of the paper similarity scores from each text model: all showing a similar clustering pattern of the three main smoothing methods. The points are colored by the most common method used in the paper, and the hulls are drawn around each method group.

5.2.3 *LLM parameters.* The text generation process of the LLMs produce a sequence of tokens based on the input text and the prompt. Model parameters such as temperature, top-p, and top-k control the randomness of the output. A higher temperature value yield more diverse outputs, while a lower value produces more conservative ones. For each Gemini extraction, we set the temperature to 1 and performed five repetitions run to assess the output stability. Among the 62 papers investigated, 31 (50%) yields a consistent number of extracted decisions across all five runs as in

573 Table 2. In 29 papers, two different counts of decisions are observed. For the remaining two papers with three versions,
 574 we investigate them separately. In the case of Ito et al. [11], variation arose because Gemini sometimes includes a
 575 collection of weather models investigated for sensitivity analysis. For Huang et al. [9], differences come from how the
 576 term “smoothing function” was sometimes extracted as the method instead of the actual method, “penalized spline” for
 577 temperature and humidity. Depending on the run, Gemini would extract “smoothing function”, “penalized spline”, or
 578 both.
 579

580 We further investigate the text similarity in the “reason” and “decision” fields for the 31 papers with the same number
 581 of decisions across runs and Table 3 tabulates the number of differences found in each pairwise comparison. Fifteen
 582 (48.4%) produce the identical text in both fields across all ten cross comparisons (5 runs yield 10 comparisons). Other
 583 showed differences in up to 11 comparisons. The primary source of variation were differing extraction amounts and
 584 failure to capture the “reason” field. For example, Kan et al. [13] has the most different (11) and the main differences are
 585 singleday lag models underestimate the cumulative effect of pollutants on mortality 2day moving
 586 average of current and previous day concentrations (lag=01)
 587 vs.
 588

589 singleday lag models underestimate the cumulative effect of pollutants on mortality 2day moving
 590 average (lag=01)
 591 and
 592 a minimum of 1 df per year was required 9 df for cardiovascular mortality
 593 vs. NA
 594

595 Table 2. Number of decisions extracted from each paper by Gemini
 596

Num. of different row number	Count
1	31
2	29
3	2

606
 607 Table 3. Number of differences in the reason and decision fields across Gemini runs for papers with consistent number of decisions
 608 across runs
 609

Num. of difference	Count	Proportion (%)
0	15	48.39
1	2	6.45
2	2	6.45
4	4	12.90
5	3	9.68
8	1	3.23
9	1	3.23
10	2	6.45
11	1	3.23

6 Discussion

- Address how sensitivity analysis is/ is not relevant
- Only prompting engineering is used to extract decisions from the literature. We expect that fine-tuning the model on statistical or domain-specific literature to yield more robust performance on the same document, though it would require substantially more training effort.
- people from the NYU-LMU workshop are interested to have code script attached as well because people can do one thing in the script but report another in the paper - it would be interesting to compare the paper and the script with some syntax extraction.
- Validation of the output:

the nature of the task: Our task involve a reasoning component in that it requires causal reasoning to identify the decisions made by the authors, and its justification/ rationale, rather than purely summarizing the text through pattern-matching.

6 References

- [1] 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>.
- [2] 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>.
- [3] 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.
- [4] 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.
- [5] 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 Barilar, 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 Glatar, Enrico Gleean, 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 Hoogeveen, Susan Holmes, Chuan-Peng Hu, Scott A. Huettel, Matthew E. Hughes, Vittorio Iacovella, Alexandru D. Iordan, Peder M. Isager, Ayse I. Isik, Andrew Jahn, Matthew R. Johnson, Tom Johnstone, Michael J. E. Joseph, Anthony C. Juliano, Joseph W. Kable, Michalis Kassinopoulos, Cemal Koba, Xiang-Zhen Kong, Timothy R. Koscik, Nuri Erkut Kucukboyaci, Brice A. Kuhl, Sebastian Kupek, Angela R. Laird, Claus Lamm, Robert Langner, Nina Lauharatanahirun, Hongmi Lee, Sangil Lee, Alexander Leemans, Andrea Leo, Elise Lesage, Flora Li, Monica Y. C. Li, Phui Cheng Lim, Evan N. Lintz, Schuyler W. Liphardt, Annabel B. Losecaat Vermeer, Bradley C. Love, Michael L. Mack, Norberto Malpica, Theo Marins, Camille Maumet, Kelsey McDonald, Joseph T. McGuire, Helena Melero, Adriana S. Méndez Leal, Benjamin Meyer, Kristin N. Meyer, Glad Mihai, Georgios D. Mitsis, Jorge Moll, Dylan M. Nielson, Gustav 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-Porxas, Emily A. Yearling, Sangsuk Yoon, Rui Yuan, Kenneth S. L. Yuen, Lei Zhang, Xu Zhang, Joshua E. Zosky, Thomas E.

- 677 Nichols, Russell A. Poldrack, and Tom Schonberg. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810):
 678 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.
 679 [6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Naacl-hlt 2019. page 4171–4186, Minneapolis, Minnesota, 06 2019. Association
 680 for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL <https://aclanthology.org/N19-1423/>.
 681 [7] 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
 682 Research Society.
 683 [8] Elliot Gould, Hannah S. Fraser, Timothy H. Parker, Shinichi Nakagawa, Simon C. Griffith, Peter A. Veski, Fiona Fidler, Daniel G. Hamilton, Robin N.
 684 Abbey-Lee, Jessica K. Abbott, Luis A. Aguirre, Carles Alcaraz, Irith Aloni, Drew Altschul, Kunal Arekar, Jeff W. Atkins, Joe Atkinson, Christopher M.
 685 Baker, Meghan Barrett, Kristian Bell, Suleiman Kehinde Bello, Iván Beltrán, Bernd J. Berauer, Michael Grant Bertram, Peter D. Billman, Charlie K.
 686 Blake, Shannon Blake, Louis Bliard, Andrea Bonisoli-Alquati, Timothée Bonnet, Camille Nina Marion Bordes, Aneesh P. H. Bose, Thomas Botterill-
 687 James, Melissa Anna Boyd, Sarah A. Boyle, Tom Bradfer-Lawrence, Jennifer Bradham, Jack A. Brand, Martin I. Brengdahl, Martin Bulla, Luc Bussière,
 688 Ettore Camerlenghi, Sara E. Campbell, Leonardo L. F. Campos, Anthony Caravaggi, Pedro Cardoso, Charles J. W. Carroll, Therese A. Catanacl,
 689 Xuan Chen, Heung Ying Janet Chik, Emily Sarah Choy, Alec Philip Christie, Angela Chuang, Amanda J. Chunco, Bethany L. Clark, Andrea Contina,
 690 Garth A. Covernton, Murray P. Cox, Kimberly A. Cressman, Marco Crotti, Connor Davidson Crouch, Pietro B. D'Amelio, Alexandra Allison
 691 de Sousa, Timm Fabian Döbert, Ralph Dobler, Adam J. Dobson, Tim S. Doherty, Szymon Marian Drobniak, Alexandra Grace Duffy, Alison B. Duncan,
 692 Robert P. Dunn, Jamie Dunning, Trishna Dutta, Luke Eberhart-Hertel, Jared Alan Elmore, Mahmoud Medhat Elsherif, Holly M. English, David C.
 693 Ensminger, Ulrich Rainer Ernst, Stephen M. Ferguson, Esteban Fernandez-Juricic, Thalita Ferreira-Arruda, John Fieberg, Elizabeth A. Finch, Evan A.
 694 Fiorenza, David N. Fisher, Amélie Fontaine, Wolfgang Forstmeier, Yoan Fourcade, Graham S. Frank, Cathryn A. Freund, Eduardo Fuentes-Lillo,
 695 Sara L. Gandy, Dustin G. Gannon, Ana I. Garcia-Cervigón, Alexis C. Garretson, Xuezen Ge, William L. Geary, Charly Géron, Marc Gilles, Antje
 696 Girndt, Daniel Glikzman, Harrison B. Goldspiel, Dylan G. E. Gomes, Megan Kate Good, Sarah C. Goslee, J. Stephen Gosnell, Eliza M. Grames, Paolo
 697 Gratton, Nicholas M. Grebe, Skye M. Greenler, Maaike Griffioen, Daniel M. Griffith, Frances J. Griffith, Jake J. Grossman, Ali Güncan, Stef Haesen,
 698 James G. Hagan, Heather A. Hager, Jonathan Philo Harris, Natasha Dean Harrison, Sarah Syedia Hasnain, Justin Chase Havird, Andrew J. Heaton,
 699 María Laura Herrera-Chaustre, Tanner J. Howard, Bin-Yan Hsu, Fabiola Iannarilli, Esperanza C. Iranzo, Erik N. K. Iverson, Saheed Olaide Jimoh,
 700 Douglas H. Johnson, Martin Johnsson, Jesse Jorna, Tommaso Jucker, Martin Jung, Ineta Kačergytė, Oliver Kaltz, Alison Ke, Clint D. Kelly, Katharine
 701 Keegan, Friedrich Wolfgang Keppeler, Alexander K. Killion, Dongmin Kim, David P. Kochan, Peter Korsten, Shan Kothari, Jonas Kuppler, Jillian M.
 702 Kusch, Małgorzata Lagisz, Kristen Marianne Lalla, Daniel J. Larkin, Courtney L. Larson, Katherine S. Lauck, M. Elise Lauterbur, Alan Law, Don-Jean
 703 Léandri-Breton, Jonas J. Lembrechts, Kiara L'Herpiniere, Eva J. P. Lievens, Daniela Oliveira de Lima, Shane Lindsay, Martin Luquet, Ross MacLeod,
 704 Kirsty H. Macphie, Kit Magellan, Magdalena M. Mair, Lisa E. Malm, Stefano Mammola, Caitlin P. Mandeville, Michael Manhart, Laura Milena
 705 Manrique-Garzon, Elina Mäntylä, Philippe Marchand, Benjamin Michael Marshall, Charles A. Martin, Dominic Andreas Martin, Jake Mitchell
 706 Martin, April Robin Martinig, Erin S. McCallum, Mark McCauley, Sabrina M. McNew, Scott J. Meiners, Thomas Merkling, Marcus Michelangeli,
 707 Maria Moiron, Bruno Moreira, Jennifer Mortensen, Benjamin Mos, Taofeek Olatunbosun Muraina, Penelope Wrenn Murphy, Luca Nelli, Petri
 708 Niemelä, Josh Nightingale, Gustav Nilsonne, Sergio Nolazco, Sabine S. Nooten, Jessie Lanterman Novotny, Agnes Birgitta Olin, Chris L. Organ,
 709 Kate L. Ostevik, Facundo Xavier Palacio, Matthieu Paquet, Darren James Parker, David J. Pascall, Valerie J. Pasquarella, John Harold Paterson, Ana
 710 Payo-Payo, Karen Marie Pedersen, Grégoire Perez, Kayla I. Perry, Patrice Pottier, Michael J. Proulx, Raphaël Proulx, Jessica L. Pruitt, Veronarinda
 711 Ramananjato, Finaritra Tolotra Randimbirison, Onja H. Razafindratsima, Diana J. Rennison, Federico Riva, Sepand Riyahi, Michael James Roast,
 712 Felipe Pereira Rocha, Dominique G. Roche, Cristian Román-Palacios, Michael S. Rosenberg, Jessica Ross, Freya E. Rowland, Deusdedith Rugemalila,
 713 Avery L. Russell, Suvi Ruuskanen, Patrick Saccone, Asaf Sadeh, Stephen M. Salazar, Kris Sales, Pablo Salmón, Alfredo Sánchez-Tójar, Leticia Pereira
 714 Santos, Francesca Santostefano, Hayden T. Schilling, Marcus Schmidt, Tim Schmoll, Adam C. Schneider, Allie E. Schrock, Julia Schroeder, Nicolas
 715 Schtickzelle, Nick L. Schultz, Drew A. Scott, Michael Peter Scroggie, Julie Teresa Shapiro, Nitika Sharma, Caroline L. Shearer, Diego Simón, Michael I.
 716 Sitvarin, Fabrício Luiz Skupien, Heather Lea Sillin, Grania Polly Smith, Jeremy A. Smith, Rahel Sollmann, Kaitlin Stack Whitney, Shannon Michael
 717 Still, Erica F. Stuber, Guy F. Sutton, Ben Swallow, Conor Claverie Taff, Elina Takola, Andrew J. Tanentzap, Rocio Tarjuelo, Richard J. Telford,
 718 Christopher J. Thawley, Hugo Thierry, Jacqueline Thomson, Svenja Tidau, Emily M. Tompkins, Claire Marie Tortorelli, Andrew Trlica, Biz R.
 719 Turnell, Lara Urban, Stijn Van de Vondel, Jessica Eva Megan van der Wal, Jens Van Eeckhoven, Francis van Oordt, K. Michelle Vanderwel, Mark C.
 720 Vanderwel, Karen J. Vanderwolf, Juliana Vélez, Diana Carolina Vergara-Florez, Brian C. Verrelli, Marcus Vinícius Vieira, Nora Villamil, Valerio
 721 Vitali, Julien Vollering, Jeffrey Walker, Xanthe J. Walker, Jonathan A. Walter, Paweł Waryszak, Ryan J. Weaver, Ronja E. M. Wedegärtner, Daniel L.
 722 Weller, and Shannon Whelan. Same data, different analysts: variation in effect sizes due to analytical decisions in ecology and evolutionary biology.
 723 *BMC Biology*, 23(1):35, 02 2025. doi: 10.1186/s12915-024-02101-x. URL <https://doi.org/10.1186/s12915-024-02101-x>.
 724 [9] Wei Huang, Jianguo Tan, Haidong Kan, Ni Zhao, Weimin Song, Guixiang Song, Guohai Chen, Lili Jiang, Cheng Jiang, Renjie Chen, and Bingheng
 725 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.
 726 scitotenv.2009.02.019. URL <https://linkinghub.elsevier.com/retrieve/pii/S004896970900165X>.
 727 [10] Nick Huntington-Klein, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pralhad Burli, Naibin Chen, Paul Grieco, Godwin
 728 Ekpe, Todd Pugatch, Martin Saavedra, and Yaniv Stopnitzky. The influence of hidden researcher decisions in applied microeconomics.
 729 *Economic Inquiry*, 59(3):944–960, 2021. doi: 10.1111/ecin.12992. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/ecin.12992>. eprint:
 730 <https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecin.12992>.

- [11] Kazuhiko Ito, William F. Christensen, Delbert J. Eatough, Ronald C. Henry, Eugene Kim, Francine Laden, Ramona Lall, Timothy V. Larson, Lucas Neas, Philip K. Hopke, and George D. Thurston. Pm source apportionment and health effects: 2. an investigation of intermethod variability in associations between source-apportioned fine particle mass and daily mortality in washington, dc. *Journal of Exposure Science & Environmental Epidemiology*, 16(4):300–310, 07 2006. doi: 10.1038/sj.jea.7500464. URL <https://www.nature.com/articles/7500464>. Publisher: Nature Publishing Group.
- [12] 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>.
- [13] 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>.
- [14] 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.
- [15] 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>.
- [16] 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/>.
- [17] 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>.
- [18] 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.
- [19] Bart Ostro, Rachel Broadwin, Shelley Green, Wen-Ying Feng, and Michael Lipsett. Fine particulate air pollution and mortality in nine california counties: Results from calfine. *Environmental Health Perspectives*, 114(1):29–33, 01 2006. doi: 10.1289/ehp.8335. URL <https://ehp.niehs.nih.gov/doi/10.1289/ehp.8335>. Publisher: Environmental Health Perspectives.
- [20] Abhraneel Sarma, Alex Kale, Michael Moon, Nathan Taback, Fanny Chevalier, Jessica Hullman, and Matthew Kay. multiverse: Multiplexing alternative data analyses in r notebooks (version 0.6.2). *OSF Preprints*, 2021. URL <https://github.com/MUCollective/multiverse>.
- [21] Marko Sarstedt, Susanne J. Adler, Christian M. Ringle, Gyeongcheol Cho, Adamantios Diamantopoulos, Heungsun Hwang, and Benjamin D. Liengaard. Same model, same data, but different outcomes: Evaluating the impact of method choices in structural equation modeling. *Journal of Product Innovation Management*, 41(6):1100–1117, 2024. doi: 10.1111/jipim.12738. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jipim.12738._eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/jipim.12738>.
- [22] R. Silberzahn, E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtry, Š. Bahnik, F. Bai, C. Bannard, E. Bonnier, R. Carlsson, F. Cheung, G. Christensen, R. Clay, M. A. Craig, A. Dalla Rosa, L. Dam, M. H. Evans, I. Flores Cervantes, N. Fong, M. Gamez-Djokic, A. Glenz, S. Gordon-McKeon, T. J. Heaton, K. Hederos, M. Heene, A. J. Hofelich Mohr, F. Högden, K. Hui, M. Johannesson, J. Kalodimos, E. Kaszubowski, D. M. Kennedy, R. Lei, T. A. Lindsay, S. Liverani, C. R. Madan, D. Molden, E. Molleman, R. D. Morey, L. B. Mulder, B. R. Nijstad, N. G. Pope, B. Pope, J. M. Prenoveau, F. Rink, E. Robusto, H. Roderique, A. Sandberg, E. Schlüter, F. D. Schönbrodt, M. F. Sherman, S. A. Sommer, K. Sotak, S. Spain, C. Spörlein, T. Stafford, L. Stefanutti, S. Tauber, J. Ullrich, M. Vianello, E.-J. Wagenmakers, M. Witkowiak, S. Yoon, and B. A. Nosek. Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, 1(3):337–356, 09 2018. doi: 10.1177/2515245917747646. URL <https://doi.org/10.1177/2515245917747646>. Publisher: SAGE Publications Inc.
- [23] Jan Simson, Fiona Draxler, Samuel Mehr, and Christoph Kern. Preventing harmful data practices by using participatory input to navigate the machine learning multiverse. CHI '25, page 1–30, New York, NY, USA, 04 2025. Association for Computing Machinery. doi: 10.1145/3706598.3713482. URL <https://dl.acm.org/doi/10.1145/3706598.3713482>.
- [24] 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>.
- [25] Hadley Wickham, Joe Cheng, and Aaron Jacobs. *ellmer: Chat with Large Language Models*, 2025. URL <https://CRAN.R-project.org/package=ellmer>. R package version 0.1.1.
- [26] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. doi: 10.48550/arXiv.1906.08237.
- [1] 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>.

- [2] 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>.
- [3] 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.
- [4] 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.
- [5] Rotem Botvinkin-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 Fed, 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 Erkut Kucukboyaci, Brice A. Kuhl, Sebastian Kupek, Angela R. Laird, Claus Lamm, Robert Langner, Nina Lauharatanahirun, Hongmi Lee, Sangil Lee, Alexander Leemans, Andrea Leo, Elise Lesage, Flora Li, Monica Y. C. Li, Phui Cheng Lim, Evan N. Lintz, Schuyler W. Liphardt, Annabel B. Losecaat Vermeer, Bradley C. Love, Michael L. Mack, Norberto Malpica, Theo Marins, Camille Maumet, Kelsey McDonald, Joseph T. McGuire, Helena Melero, Adriana S. Méndez Leal, Benjamin Meyer, Kristin N. Meyer, Glad Mihai, Georgios D. Mitsis, Jorge Moll, Dylan M. Nielson, Gustav Nilsonne, Michael P. Notter, Emanuele Olivetti, Adrian I. Onicas, Paolo Papale, Kaustubh R. Patil, Jonathan E. Peele, 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. Nichols, Russell A. Poldrack, and Tom Schonberg. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810):84–88, 06 2020. doi: 10.1038/s41586-020-2314-9. URL <https://www.nature.com/articles/s41586-020-2314-9>. Publisher: Nature Publishing Group.
- [6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Naacl-hlt 2019. page 4171–4186, Minneapolis, Minnesota, 06 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL <https://aclanthology.org/N19-1423>.
- [7] Andrew Gelman and Eric Loken. The statistical crisis in science. *American Scientist*, 102(6):460–465, 12 2014. URL <https://www.proquest.com/docview/1616141998/abstract/5E050DCE82414037PQ/1>. Num Pages: 6 Place: Research Triangle Park, United States Publisher: Sigma XI-The Scientific Research Society.
- [8] Elliot Gould, Hannah S. Fraser, Timothy H. Parker, Shinichi Nakagawa, Simon C. Griffith, Peter A. Veski, Fiona Fidler, Daniel G. Hamilton, Robin N. Abbey-Lee, Jessica K. Abbott, Luis A. Aguirre, Carles Alcaraz, Irith Aloni, Drew Altschul, Kunal Arekar, Jeff W. Atkins, Joe Atkinson, Christopher M. Baker, Meghan Barrett, Kristian Bell, Suleiman Kehinde Bello, Iván Beltrán, Bernd J. Berauer, Michael Grant Bertram, Peter D. Billman, Charlie K. Blake, Shannon Blake, Louis Bliard, Andrea Bonisoli-Alquati, Timothée Bonnet, Camille Nina Marion Bordes, Aneesh P. H. Bose, Thomas Botterill-James, Melissa Anna Boyd, Sarah A. Boyle, Tom Bradfer-Lawrence, Jennifer Bradham, Jack A. Brand, Martin I. Brengdahl, Martin Bulla, Luc Bussière, Ettore Camerlenghi, Sara E. Campbell, Leonardo L. F. Campos, Anthony Caravaggi, Pedro Cardoso, Charles J. W. Carroll, Therese A. Catanach, Xuan Chen, Heung Ying Janet Chik, Emily Sarah Choy, Alec Philip Christie, Angela Chuang, Amanda J. Chunco, Bethany L. Clark, Andrea Contina, Garth A. Covernton, Murray P. Cox, Kimberly A. Cressman, Marco Crotti, Connor Davidson Crouch, Pietro B. D'Amelio, Alexandra Allison de Sousa, Timm Fabian Döbert, Ralph Dobler, Adam J. Dobson, Tim S. Doherty, Szymon Marian Drobniaik, Alexandra Grace Duffy, Alison B. Duncan, Robert P. Dunn, Jamie Dunning, Trishna Dutta, Luke Eberhart-Hertel, Jared Alan Elmore, Mahmoud Medhat Elsherif, Holly M. English, David C. Ensminger, Ulrich Rainer Ernst, Stephen M. Ferguson, Esteban Fernandez-Juricic, Thalita Ferreira-Arruda, John Fieberg, Elizabeth A. Finch, Evan A. Fiorenza, David N. Fisher, Amélie Fontaine, Wolfgang Forstmeier, Yoan Fourcade, Graham S. Frank, Cathryn A. Freund, Eduardo Fuentes-Lillo, Sara L. Gandy, Dustin G. Gannon, Ana I. García-Cervigón, Alexis C. Garretson, Xuezen Ge, William L. Geary, Charly Géron, Marc Gilles, Antje Girndt, Daniel Gliksman, Harrison B. Goldspiel, Dylan G. E. Gomes, Megan Kate Good, Sarah C. Goslee, J. Stephen Gosnell, Eliza M. Grames, Paolo Gratton, Nicholas M. Grebe, Skye M. Greenler, Maaike Griffioen, Daniel M. Griffith, Frances J. Griffith, Jake J. Grossman, Ali Güncan, Stef Haesen, James G. Hagan, Heather A. Hager, Jonathan Philo Harris, Natasha Dean Harrison, Sarah Syedia Hasnain, Justin Chase Havird, Andrew J. Heaton,

- 833 María Laura Herrera-Chaustre, Tanner J. Howard, Bin-Yan Hsu, Fabiola Iannarilli, Esperanza C. Iranzo, Erik N. K. Iverson, Saheed Olaide Jimoh,
 834 Douglas H. Johnson, Martin Johnsson, Jesse Jorna, Tommaso Jucker, Martin Jung, Ineta Kačergytė, Oliver Kaltz, Alison Ke, Clint D. Kelly, Katharine
 835 Keogan, Friedrich Wolfgang Keppeler, Alexander K. Killion, Dongmin Kim, David P. Kochan, Peter Korsten, Shan Kothari, Jonas Kuppler, Jillian M.
 836 Kusch, Małgorzata Lagisz, Kristen Marianne Lalla, Daniel J. Larkin, Courtney L. Larson, Katherine S. Lauck, M. Elise Lauterbur, Alan Law, Don-Jean
 837 Léandri-Breton, Jonas J. Lembrechts, Kiara L'Herpiniere, Eva J. P. Lievens, Daniela Oliveira de Lima, Shane Lindsay, Martin Luquet, Ross MacLeod,
 838 Kirsty H. Macphie, Kit Magellan, Magdalena M. Mair, Lisa E. Malm, Stefano Mammola, Caitlin P. Mandeville, Michael Manhart, Laura Milena
 839 Manrique-Garzon, Elina Mäntylä, Philippe Marchand, Benjamin Michael Marshall, Charles A. Martin, Dominic Andreas Martin, Jake Mitchell
 840 Martin, April Robin Marting, Erin S. McCallum, Mark McCauley, Sabrina M. McNew, Scott J. Meiners, Thomas Merkling, Marcus Michelangeli,
 841 Maria Moiron, Bruno Moreira, Jennifer Mortensen, Benjamin Mos, Tacfeek Olatunbosun Muraina, Penelope Wrenn Murphy, Luca Nelli, Petri
 842 Niemelä, Josh Nightingale, Gustav Nilsonne, Sergio Nolazco, Sabine S. Nooten, Jessie Lanterman Novotny, Agnes Birgitta Olin, Chris L. Organ,
 843 Kate L. Ostevik, Facundo Xavier Palacio, Matthieu Paquet, Darren James Parker, David J. Pascall, Valerie J. Pasquarella, John Harold Paterson, Ana
 844 Payo-Payo, Karen Marie Pedersen, Grégoire Perez, Kayla I. Perry, Patrice Pottier, Michael J. Proulx, Raphaël Proulx, Jessica L. Pruitt, Veronarindra
 845 Ramananjato, Finaritra Tolotra Randimbison, Onja H. Razafindratsima, Diana J. Rennison, Federico Riva, Sepand Riyahi, Michael James Roast,
 846 Felipe Pereira Rocha, Dominique G. Roche, Cristian Román-Palacios, Michael S. Rosenberg, Jessica Ross, Freya E. Rowland, Deusdedit Rugemalila,
 847 Avery L. Russell, Suvi Ruuskanen, Patrick Saccone, Asaf Sadeh, Stephen M. Salazar, Kris Sales, Pablo Salmón, Alfredo Sánchez-Tójar, Leticia Pereira
 848 Santos, Francesca Santostefano, Hayden T. Schilling, Marcus Schmidt, Tim Schmoll, Adam C. Schneider, Allie E. Schrock, Julia Schroeder, Nicolas
 849 Schtickzelle, Nick L. Schultz, Drew A. Scott, Michael Peter Scroggie, Julie Teresa Shapiro, Nitika Sharma, Caroline L. Shearer, Diego Simón, Michael I.
 850 Sitvarin, Fabrício Luiz Skupien, Heather Lea Slinn, Grania Polly Smith, Jeremy A. Smith, Rahel Sollmann, Kaitlin Stack Whitney, Shannon Michael
 851 Still, Erica F. Stuber, Guy F. Sutton, Ben Swallow, Conor Claverie Taff, Elina Takola, Andrew J. Tanentzap, Rocío Tarjuelo, Richard J. Telford,
 852 Christopher J. Thawley, Hugo Thierry, Jacqueline Thomson, Svenja Tidau, Emily M. Tompkins, Claire Marie Tortorelli, Andrew Trlica, Biz R.
 853 Turnell, Lara Urban, Stijn Van de Vondel, Jessica Eva Megan van der Wal, Jens Van Eeckhoven, Francis van Oordt, K. Michelle Vanderwel, Mark C.
 854 Vanderwel, Karen J. Vanderwolf, Juliana Vélez, Diana Carolina Vergara-Florez, Brian C. Verrelli, Marcus Vinícius Vieira, Nora Villamil, Valerio
 855 Vitali, Julien Vollering, Jeffrey Walker, Xanthe J. Walker, Jonathan A. Walter, Paweł Waryszak, Ryan J. Weaver, Ronja E. M. Wedegärtner, Daniel L.
 856 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>.
- [9] Wei Huang, Jianguo Tan, Haidong Kan, Ni Zhao, Weimin Song, Guixiang Song, Guohai Chen, Lili Jiang, Cheng Jiang, Renjie Chen, and Bingheng
 857 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.
 858 scitotenv.2009.02.019. URL <https://linkinghub.elsevier.com/retrieve/pii/S004896970900165X>.
- [10] Nick Huntington-Klein, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pralhad Burli, Naibin Chen, Paul Grieco, Godwin
 859 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:
 860 <https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecin.12992>.
- [11] Kazuhiko Ito, William F. Christensen, Delbert J. Eatough, Ronald C. Henry, Eugene Kim, Francine Laden, Ramona Lall, Timothy V. Larson, Lucas
 861 Neas, Philip K. Hopke, and George D. Thurston. Pm source apportionment and health effects: 2. an investigation of intermethod variability in
 862 associations between source-apportioned fine particle mass and daily mortality in washington, dc. *Journal of Exposure Science & Environmental
 Epidemiology*, 16(4):300–310, 07 2006. doi: 10.1038/sj.jea.7500464. URL <https://www.nature.com/articles/7500464>. Publisher: Nature Publishing
 863 Group.
- [12] Alex Kale, Matthew Kay, and Jessica Hullman. Decision-making under uncertainty in research synthesis: Designing for the garden of forking
 864 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>.
- [13] Haidong Kan, Stephanie J. London, Guohai Chen, Yunhui Zhang, Guixiang Song, Naiqing Zhao, Lili Jiang, and Bingheng Chen. Differentiating the
 865 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.
 866 2006.12.001. URL <https://www.sciencedirect.com/science/article/pii/S0160412006002108>.
- [14] Klea Katsouyanni, Giota Touloumi, Evangelia Samoli, Alexandros Gryparis, Alain Le Tertre, Yannis Monopolis, Giuseppe Rossi, Denis Zmirou,
 867 Ferran Ballester, Azedine Boumghar, Hugh Ross Anderson, Bogdan Wojtyniak, Anna Paldy, Rony Braunstein, Juha Pekkanen, Christian Schindler,
 868 and Joel Schwartz. Confounding and effect modification in the short-term effects of ambient particles on total mortality: Results from 29 european
 869 cities within the apeha2 project. *Epidemiology*, 12(5):521, 09 2001. URL https://journals.lww.com/epidem/fulltext/2001/09000/confounding_and_effect_modification_in_the_11.aspx.
- [15] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical
 870 language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 02 2020. doi: 10.1093/bioinformatics/btz682. URL
 871 <https://academic.oup.com/bioinformatics/article/36/4/1234/5566506>.
- [16] Jiali Liu, Nadia Boukhelifa, and James R. Eagan. Understanding the Role of Alternatives in Data Analysis Practices. *IEEE Transactions on Visualization
 872 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/>.
- [17] Yang Liu, Tim Althoff, and Jeffrey Heer. Paths explored, paths omitted, paths obscured: Decision points & selective reporting in end-to-end
 873 data analysis. CHI '20, page 1–14, New York, NY, USA, 04 2020. Association for Computing Machinery. doi: 10.1145/3313831.3376533. URL
 874 Manuscript submitted to ACM

- 885 <https://dl.acm.org/doi/10.1145/3313831.3376533>.

886 [18] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov.
887 Roberta: A robustly optimized bert pretraining approach. doi: 10.48550/arXiv.1907.11692.

888 [19] Bart Ostro, Rachel Broadwin, Shelley Green, Wen-Ying Feng, and Michael Lipsett. Fine particulate air pollution and mortality in nine California
889 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.

890 [20] Abhraneel Sarma, Alex Kale, Michael Moon, Nathan Taback, Fanny Chevalier, Jessica Hullman, and Matthew Kay. multiverse: Multiplexing
891 alternative data analyses in r notebooks (version 0.6.2). *OSF Preprints*, 2021. URL <https://github.com/MUCollective/multiverse>.

892 [21] Marko Sarstedt, Susanne J. Adler, Christian M. Ringle, Gyeongcheol Cho, Adamantios Diamantopoulos, Heungsun Hwang, and Benjamin D.
893 Liengard. Same model, same data, but different outcomes: Evaluating the impact of method choices in structural equation modeling. *Journal of
894 Product Innovation Management*, 41(6):1100–1117, 2024. doi: 10.1111/jpim.12738. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/jpim.12738>.
895 _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/jpim.12738>.

896 [22] R. Silberzahn, E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtry, Š. Bahnik, F. Bai, C. Bannard, E. Bonnier, R. Carlsson, F. Cheung,
897 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,
898 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,
899 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,
900 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,
901 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
902 transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, 1(3):337–356, 09 2018. doi:
903 10.1177/2515245917747646. URL <https://doi.org/10.1177/2515245917747646>. Publisher: SAGE Publications Inc.

904 [23] Jan Simson, Fiona Draxler, Samuel Mehr, and Christoph Kern. Preventing harmful data practices by using participatory input to navigate the
905 machine learning multiverse. CHI ’25, page 1–30, New York, NY, USA, 04 2025. Association for Computing Machinery. doi: 10.1145/3706598.3713482.
906 URL <https://dl.acm.org/doi/10.1145/3706598.3713482>.

907 [24] 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>.

908 [25] Hadley Wickham, Joe Cheng, and Aaron Jacobs. *ellmer: Chat with Large Language Models*, 2025. URL <https://CRAN.R-project.org/package=ellmer>.
909 R package version 0.1.1.

910 [26] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for
911 language understanding. doi: 10.48550/arXiv.1906.08237.

912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936 Manuscript submitted to ACM