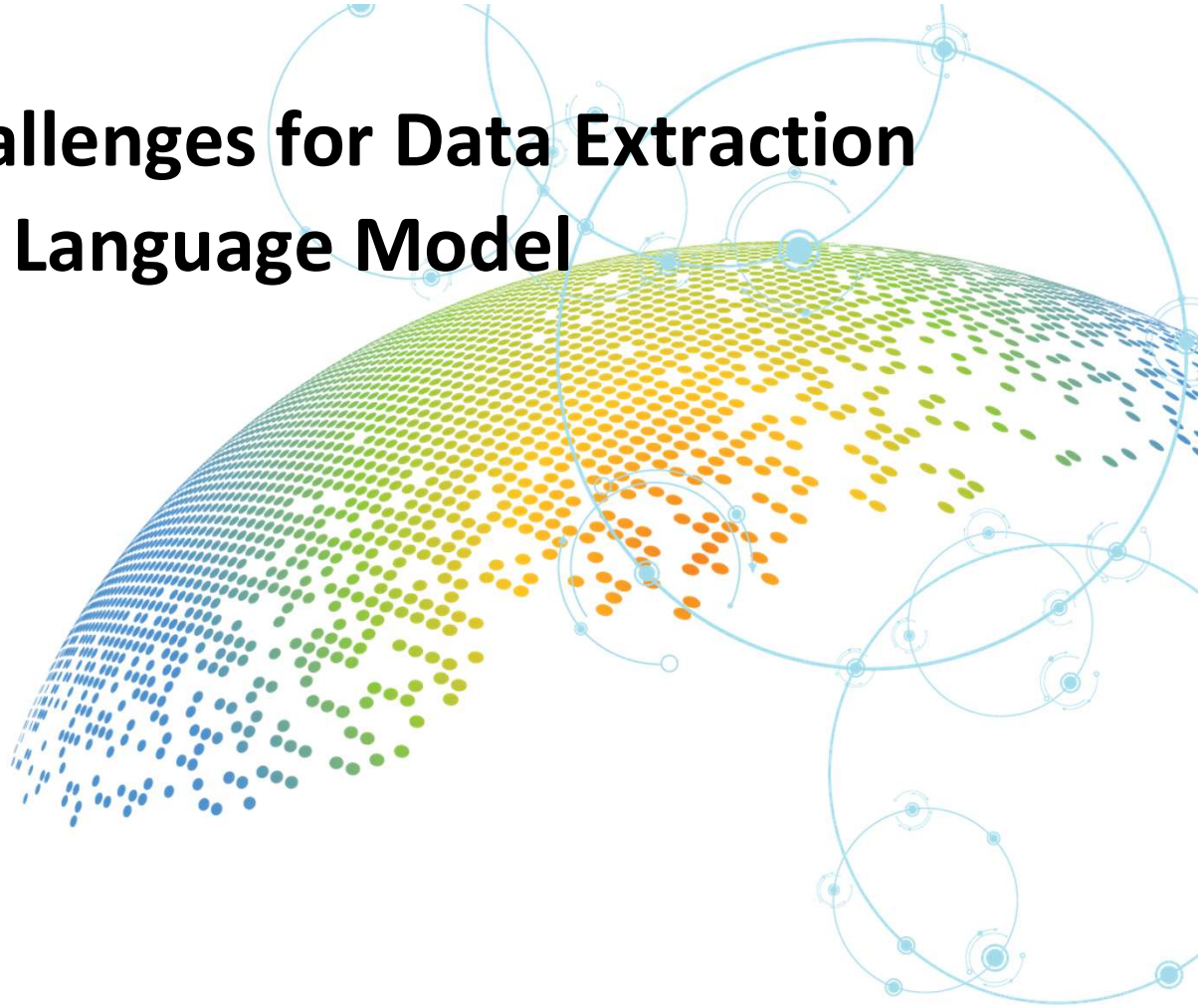
Opportunities and Challenges for Data Extraction with a Large Language Model



Gerald Gartlehner

Cochrane Austria

RTI International

March 12, 2025

RTI INTERNATIONAL

Declaration of Conflict of Interest

I have no actual or potential conflict of interest in relation to this presentation.

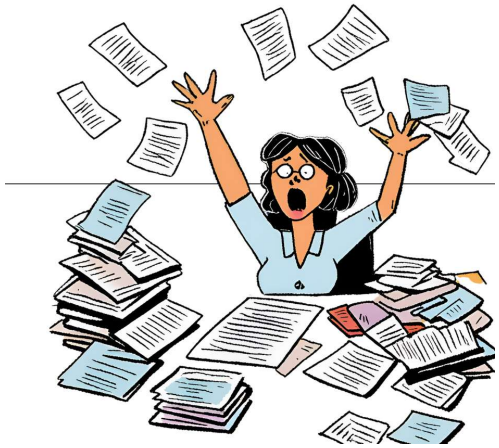
The research was funded by the RTI International Innovation Office and the US Agency for Healthcare Research and Quality (AHRQ).

Overview

- Data extraction for evidence synthesis
- Use of Large Language Models (LLMs) in data extraction
- Research findings on LLM-based data extraction
- Practical guidance for LLM-assisted data extraction

Data Extraction

- The process of transcribing data from primary studies into standardized tables.
- **Complexity ranges** from simple copying and pasting to performing transformations or calculations.
- Data extraction is often **time-consuming, costly, tedious, and prone to errors**.
- Up to 50% of studies included in systematic reviews had at least one data extraction error.



Created by Canva (2025)

Mathes T et al. BMC Med Res Methodol. 2017:17(1):152.

Data Extraction - Errors

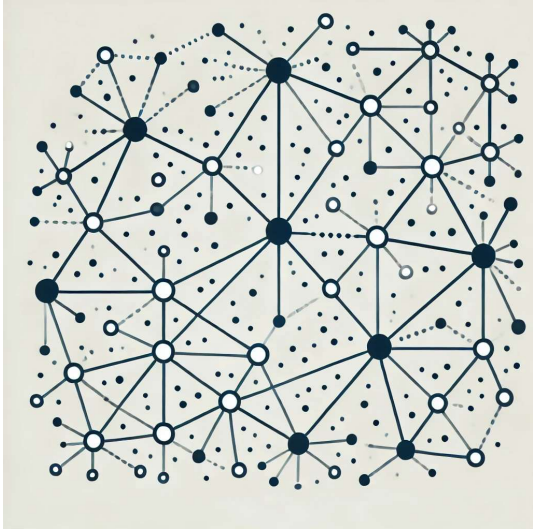
Data Extraction Method	Proportion of Errors (ranges)
Single-reviewer extraction	34% - 36%
Single-reviewer extraction with verification by second reviewer	16% - 24%
Dual, independent extraction by two reviewers	14% -16%

MECIR: Dual, independent extraction is **mandatory** for outcome data, and **highly desirable** for study characteristics

Buscemi N et al. J Clin Epidemiol. 2006: 59(7):697-703.
Li T et al. J Clin Epidemiol. 2019. 115:77-89.
Horton J et al. J Clin Epidemiol. 2010. 63(3):289-98.
Tang L et al. medRxiv. 2024: https://doi.org/10.1101/2023.10.16.23297056.

Use of Artificial Intelligence for Data Extraction

- Previous methods primarily employed **natural language processing**, using statistical models (e.g., support vector machines, Bayesian models).
- Tools typically require extensive **labeled training datasets** and often failed to achieve sufficient accuracy.
- Most tools encounter difficulties when extracting data from **tables and figures** within PDFs.
- Accuracy of data extraction from full texts ranges from **69% to 90%.**



Schmidt L et al. F1000 Research. 2023.10: 401

Generative Large Language Models (LLMs)

- Generative LLMs are AI systems trained on extensive datasets to **predict subsequent tokens** (words, subwords, or characters) in a sequence of words.
- They primarily utilize **transformer-based** deep learning architectures to generate contextually relevant responses.
- LLMs facilitate **zero-shot applications** in data extraction, without additional training or programming.
- LLMs had remarkable **gains in speed and the capacity** to process large volumes of text.

Image by Tumisu from Pixabay

Model Capacities

LLM-Model	Context Window (tokens)	Maximum Length of PDF Document*
ChatGPT 4 standard	8,192	~11 pages
DeepSeek	64,000	~85 pages
ChatGPT 4o	128,000	~170 pages
Claude 3.5 Sonnet	200,000	~265 pages
Gemini 1.5 Pro	1,000,000	~1300 pages

*Assumes 1 token = 0.75 words; 1 single spaced page = 500 words

PW0

Evaluations of LLMs for Data Extraction

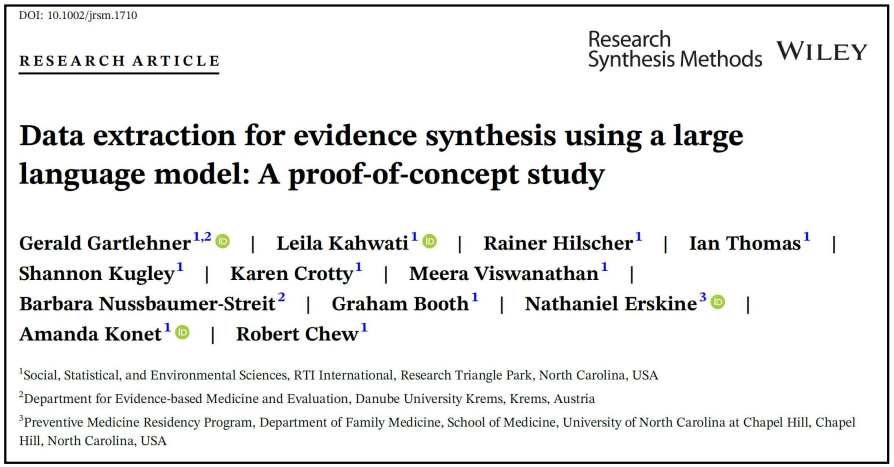
- Initial evaluations of LLMs for data extraction from full-text PDFs reported variable accuracy (**ranging from 72% to 100%)** when compared to human reference standards.
- Limitations:
  - Existing studies relied on controlled experimental conditions and pre-existing review datasets as benchmarks, which may limit generalizability to real-world scenarios.
  - Evaluations primarily focused on fully automated approaches without human involvement.

Motzfeldt Jensen M et al. PLoS One. 2025:20(1):e0313401.
JiayiLiu M et al. Int J Surg. 2025:10.1097.
Khan M et al. J Am Med Inform Assoc. 2025:1-10.
Gartlehner G et al. Res Synth Methods. 2024:15(4):576-89..
Khraisha Q et al. Res Synth Methods. 2024:15(4):616-26.
Konet A et al. Res Synth Methods. 2024:15(5):818-24.
Panayi A et al. Syst Rev. 2023:12(1):187.
Mahmoudi H et al. Preprint available at SSRN 4797024.

PW0 Diese beiden Referenzen habe ich in PubMed/Scholar nicht gefunden.
Petra Wellemsen, 2025-03-10T09:20:34.057

Study Design

- **Validation study** assessing Claude performance in data extraction.
- Reference standard: **Enhanced manual data extraction by humans.**
- Convenience sample of **10 open-access journal publications** of RCTs provided as PDFs.
- **16 data elements** including study and population characteristics, outcomes data, participant flow, etc.



DOI: 10.1002/jrsm.1710

RESEARCH ARTICLE

Research Synthesis Methods WILEY

Data extraction for evidence synthesis using a large language model: A proof-of-concept study

Gerald Gartlehner[1,2] | **Leila Kahwati**[1] | **Rainer Hilscher**[1] | **Ian Thomas**[1] | **Shannon Kugley**[1] | **Karen Crotty**[1] | **Meera Viswanathan**[1] | **Barbara Nussbaumer-Streit**[2] | **Graham Booth**[1] | **Nathaniel Erskine**[3] | **Amanda Konet**[1] | **Robert Chew**[1]

[1]Social, Statistical, and Environmental Sciences, RTI International, Research Triangle Park, North Carolina, USA

[2]Department for Evidence-based Medicine and Evaluation, Danube University Krems, Krems, Austria

[3]Preventive Medicine Residency Program, Department of Family Medicine, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

Data Sources

Secukinumab is superior to ustekinumab in clearing skin of subjects with moderate to severe plaque psoriasis: CLEAR, a randomized controlled trial

Diamant Thaçi, MD,[a] Andrew Blauvelt, MD, MBA,[b] Kristian Reich, MD,[c] Tsen-Fang Tsai, MD,[d] Francisco Vanaclocha, MD,[e] Külli Kingo, MD, PhD,[f] Michael Ziv, MD, BSc,[g] Andreas Pinter, MD,[h] Sophie Hugot, MSc,[i] Ruquan You, MSc,[j] and Marina Milutinovic, MD[i]

Lübeck, Göttingen, and Frankfurt, Germany; Portland, Oregon; Taipei, Taiwan; Madrid, Spain; Tartu, Estonia; Afula, Israel; Basel, Switzerland; and Shanghai, China

Background: Secukinumab, a fully human anti-interleukin-17A monoclonal antibody, has shown superior efficacy to etanercept with similar safety in moderate to severe plaque psoriasis (FIXTURE study).

Objective: We sought to directly compare efficacy and safety of secukinumab versus ustekinumab.

Methods: In this 52-week, double-blind study (NCT02074982), 676 subjects were randomized 1:1 to subcutaneous injection of secukinumab 300 mg or ustekinumab per label. Primary end point was 90% or more improvement from baseline Psoriasis Area and Severity Index (PASI) score (PASI 90) at week 16.

Results: Secukinumab (79.0%) was superior to ustekinumab (57.6%) as assessed by PASI 90 response at week 16 ($P < .0001$). The 100% improvement from baseline PASI score at week 16 was also significantly greater with secukinumab (44.3%) than ustekinumab (28.4%) ($P < .0001$). The 75% or more improvement from baseline PASI score at week 4 was superior for secukinumab (50.0%) versus ustekinumab (20.6%) ($P < .0001$). Percentage of subjects with the Dermatology Life Quality Index score 0/1 (week 16) was significantly higher with secukinumab (71.9%) than ustekinumab (57.4%) ($P < .0001$). The safety profile of secukinumab was comparable with ustekinumab and consistent with pivotal phase III secukinumab studies.

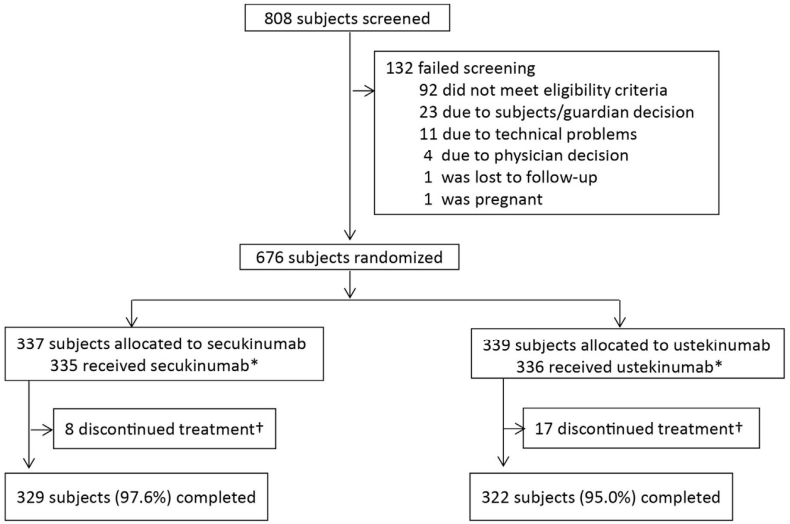


Table I. Baseline demographic and clinical characteristics

Characteristic	Secukinumab 300 mg (n = 337)	Ustekinumab (n = 339)
Age, y	45.2 ± 13.96	44.6 ± 13.67
Male gender	229 (68.0)	252 (74.3)
Race		
Caucasian	299 (88.7)	288 (85.0)
Other	38 (11.3)	51 (15.0)
Weight, kg	87.4 ± 19.95	87.2 ± 22.11
BMI, kg/m^2	29.1 ± 5.87	29.0 ± 6.69
Time since psoriasis diagnosis, y	19.6 ± 12.90	16.1 ± 11.24
PASI score	21.7 ± 8.50	21.5 ± 8.07
Body surface area involved, %	32.6 ± 17.78	32.0 ± 16.80
IGA mod 2011 score		
4 (Severe disease)*	130 (38.6)	125 (36.9)
Psoriatic arthritis reported	69 (20.5)	54 (15.9)
Previous systemic psoriasis treatment		
Any	225 (66.8)	231 (68.1)
Conventional agent†	218 (64.7)	223 (65.8)
Biologic agent	48 (14.2)	44 (13.0)
Failed biologic agent	36 (10.7)	34 (10.0)

Performance

- Overall **accuracy: 96.3%.**
- **6 errors** out 160 data elements
- Claude detected several minor **errors in human data extractions**.

1 major error

Fabricated data for 1 treatment group (hallucination)

1 minor error

Rounding error of standard deviation

Missed data

In 4 instances

Gartlehner G et al. Res Synth Methods. 2024:15(4):576-89.

Output

Trial registry number: NCT02905006

Female participants:
Placebo: 17 (40.5%)
Bimekizumab 64 mg: 19 (48.7%)
Bimekizumab 160 mg: 11 (25.6%)

Characteristic	Placebo Q4W, N = 42	BKZ 64 mg Q4W, N = 39	BKZ 160 mg Q4W, N = 43
Age, years, mean (SD)	46.7 (12.3)	44.2 (13.8)	43.4 (12.4)
Male sex, n (%)	25 (59.5)	20 (51.3)	32 (74.4)
Weight, kg, mean (SD)	88.8 (21.2)	86.6 (19.2)	91.6 (24.6)

How did you calculate female participants?

The **total number of participants** in each treatment group is provided in the table "Baseline demographics and disease characteristics". The percentage of male participants in each group is also provided in the table.
To calculate the number of female participants, **I subtracted the number of male participants from the total number** of participants in each group. To calculate the percentage of female participants in each group, **I divided the number of female participants by the total number of participants** and multiplied by 100.

Study within Reviews (SWAR)

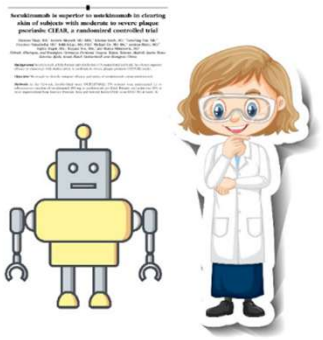
- Six **use cases under "real-world"** circumstances of systematic reviews.

Traditional human-only data extraction

vs.

- Concordance
- Accuracy
- Time required

Semi-automated data extraction replacing one human



Study Design: Prospective Parallel Group Study

Studies included in systematic review

Review team defines data elements that need to be extracted

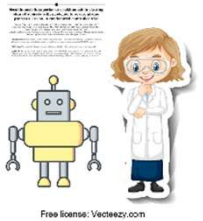
Team 1 (part of review team)
Human-only data extraction

- Data extraction by one investigator
- Validation of completeness and correctness by a second investigator
- Resolution of discrepancies

B L I N D I N G

Team 2 (part of review team)
Semi-automated data extraction

- Prompt engineering
- Data extraction by Claude
- Validation of completeness and correctness by a human investigator
- Resolution of discrepancies



Extracted data

Extracted data

Adjudication Team
Comparison and resolution of discrepancies by independent and blinded adjudicators

Study Design: Prospective Parallel Group Study

Studies included in systematic review

Review team defines data elements that need to be extracted

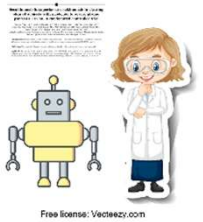
Team 1 (part of review team)
Human-only data extraction

- Data extraction by one investigator
- Validation of completeness and correctness by a second investigator
- Resolution of discrepancies

B L I N D I N G

Team 2 (part of review team)
Semi-automated data extraction

- Prompt engineering
- Data extraction by Claude
- Validation of completeness and correctness by a human investigator
- Resolution of discrepancies



Extracted data

Adjudication Team
Comparison and resolution of discrepancies by independent and blinded adjudicators

Extracted data

Study Design: Prospective Parallel Group Study

Studies included in systematic review

Review team defines data elements that need to be extracted

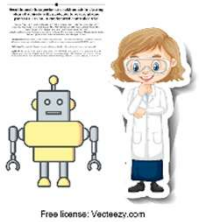
Team 1 (part of review team)
Human-only data extraction

- Data extraction by one investigator
- Validation of completeness and correctness by a second investigator
- Resolution of discrepancies

B L I N D I N G

Team 2 (part of review team)
Semi-automated data extraction

- Prompt engineering
- Data extraction by Claude
- Validation of completeness and correctness by a human investigator
- Resolution of discrepancies



Extracted data

Extracted data

Adjudication Team
Comparison and resolution of discrepancies by independent and blinded adjudicators

Tasks of Adjudicating Team

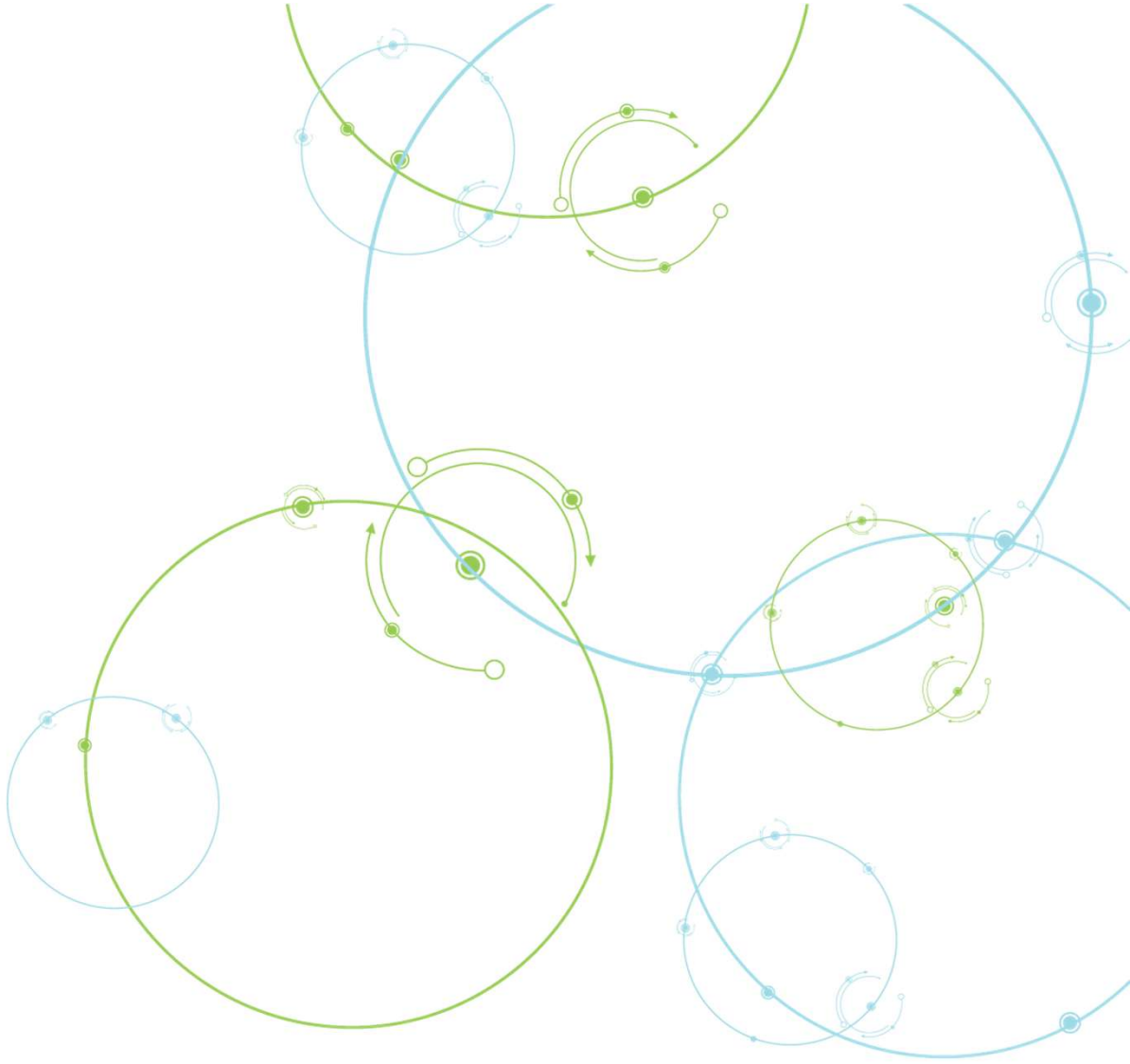
- Evaluation of the **concordance of extracted data** and **classification of errors**.
- For any **discrepancies in extracted data**, adjudicators checked the journal publications.
- In cases where data extractions by humans were incorrect, they **revised reference standard**.

Concordance	The **factual congruence** of extracted data items, even if there are variations in style, presentation, or length between the two data extractions.

Severity of Errors

Error	Definitions
Major error	This error significantly compromises the accuracy of the data, and, if uncorrected, could lead to erroneous conclusions.
Minor error	This error is less severe than a major error and may or may not impact interpretation of the existing data.
Inconsequential difference	This difference most likely would not impact the interpretation of the data

Findings



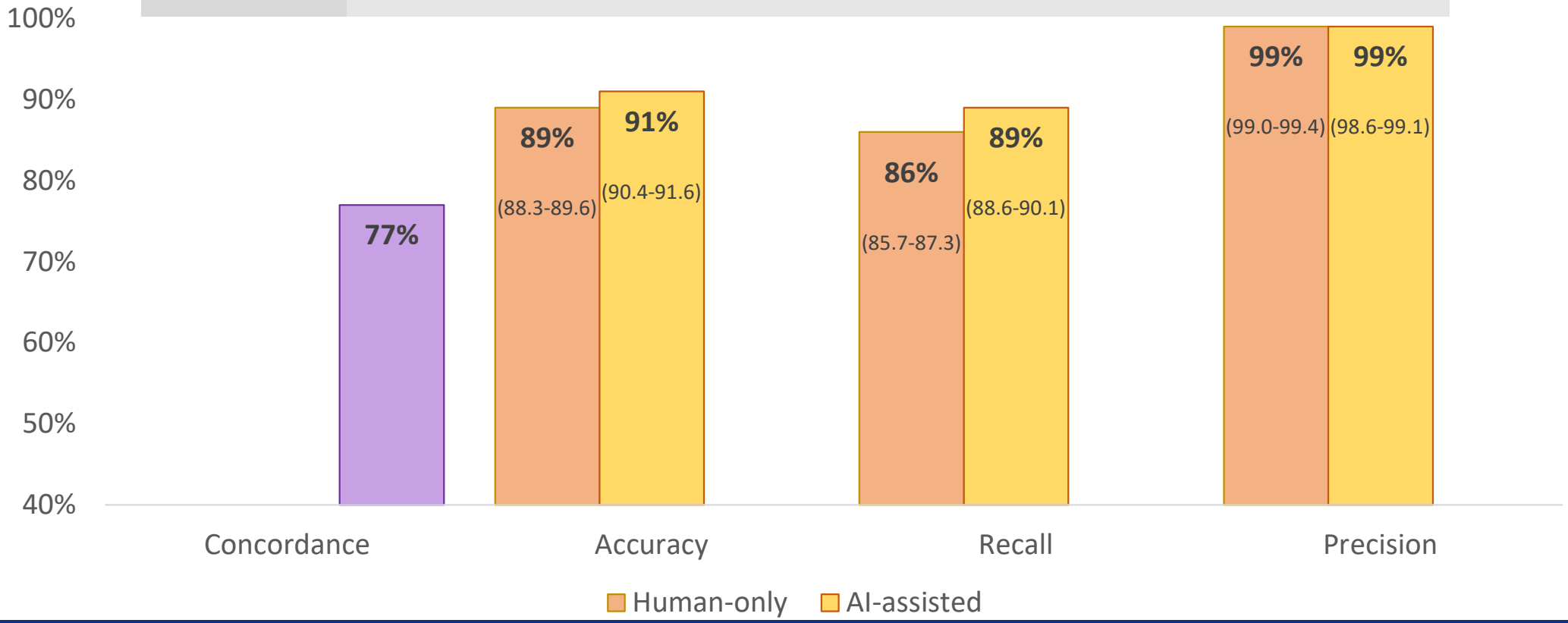
Characteristics of Reviews

Topic	k Studies	n Data Items
Implementation Strategies to Prevent Mental Health Disorders in Children/Adolescents	11*	891
Interventions to Improve Care of Bereaved Persons	20*	1,337
Prevention of Tobacco Use in Children and Adolescent	7*	292
Prevention of Hospital-acquired Infections	10*	1,797
Peripheral Nerve Blocks for Postoperative Pain Management in Cardiothoracic Surgery	8	1,759
Behavioral Counseling Interventions to Prevent Cardiovascular Disease in Adults:	7	3,265
Total	**k=63**	**n=9,341**

* Includes RCTs and NRSIs

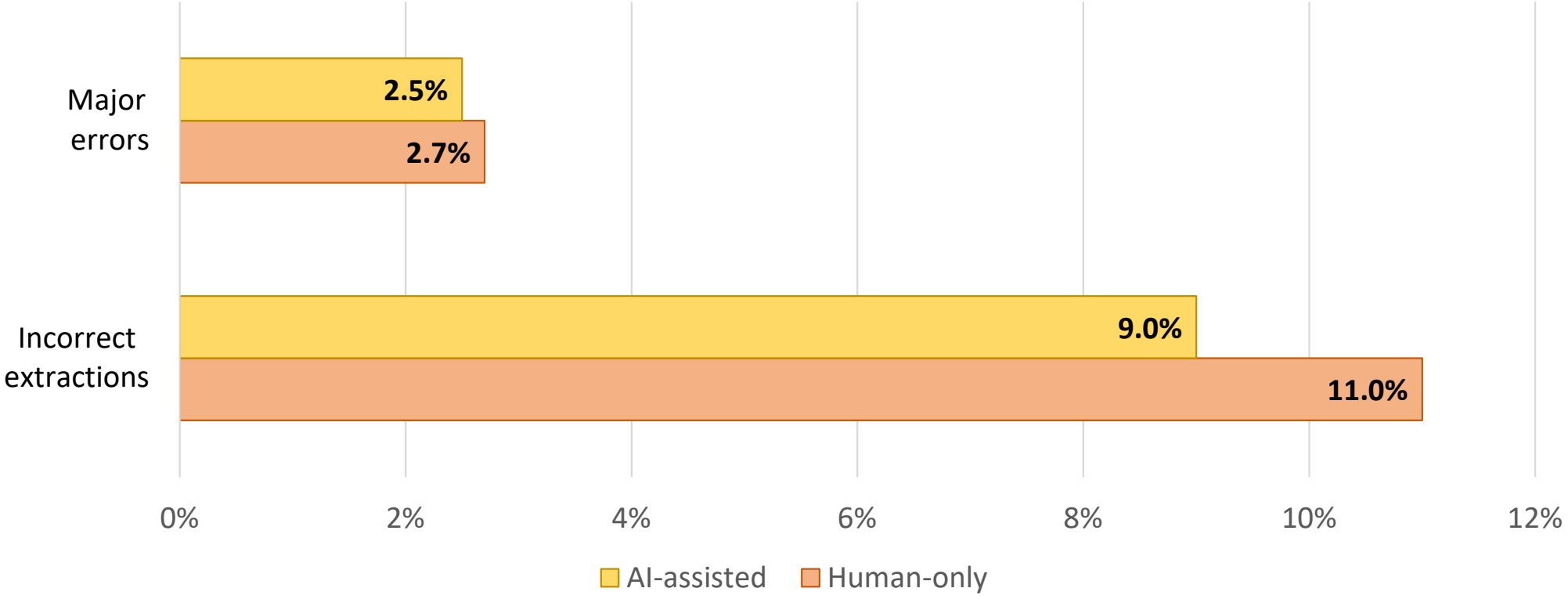
Performance Metrics

Recall	The ability of a data extraction approach to correctly extract available data
Precision	The correctness of extracted data items



Incorrect Extractions and Major Errors

Major error	This error significantly compromises the accuracy of the data, and, if uncorrected, could lead to erroneous conclusions.

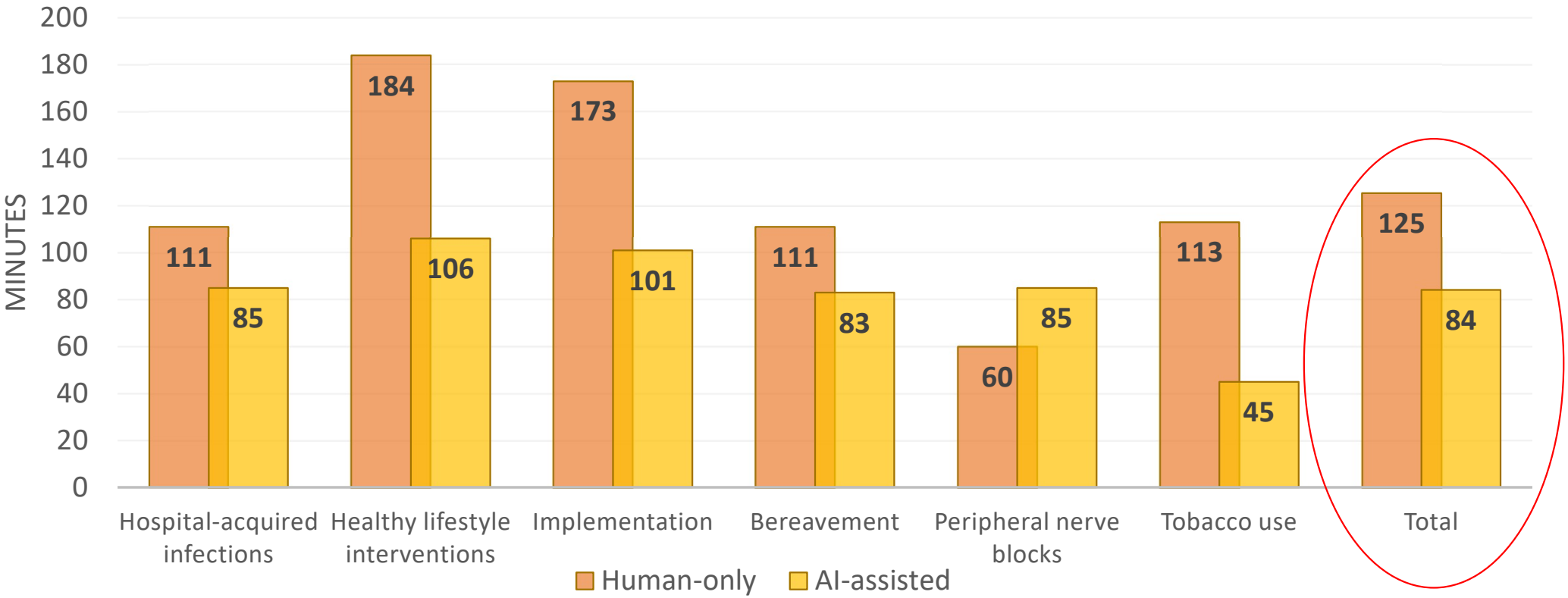


Types of Errors

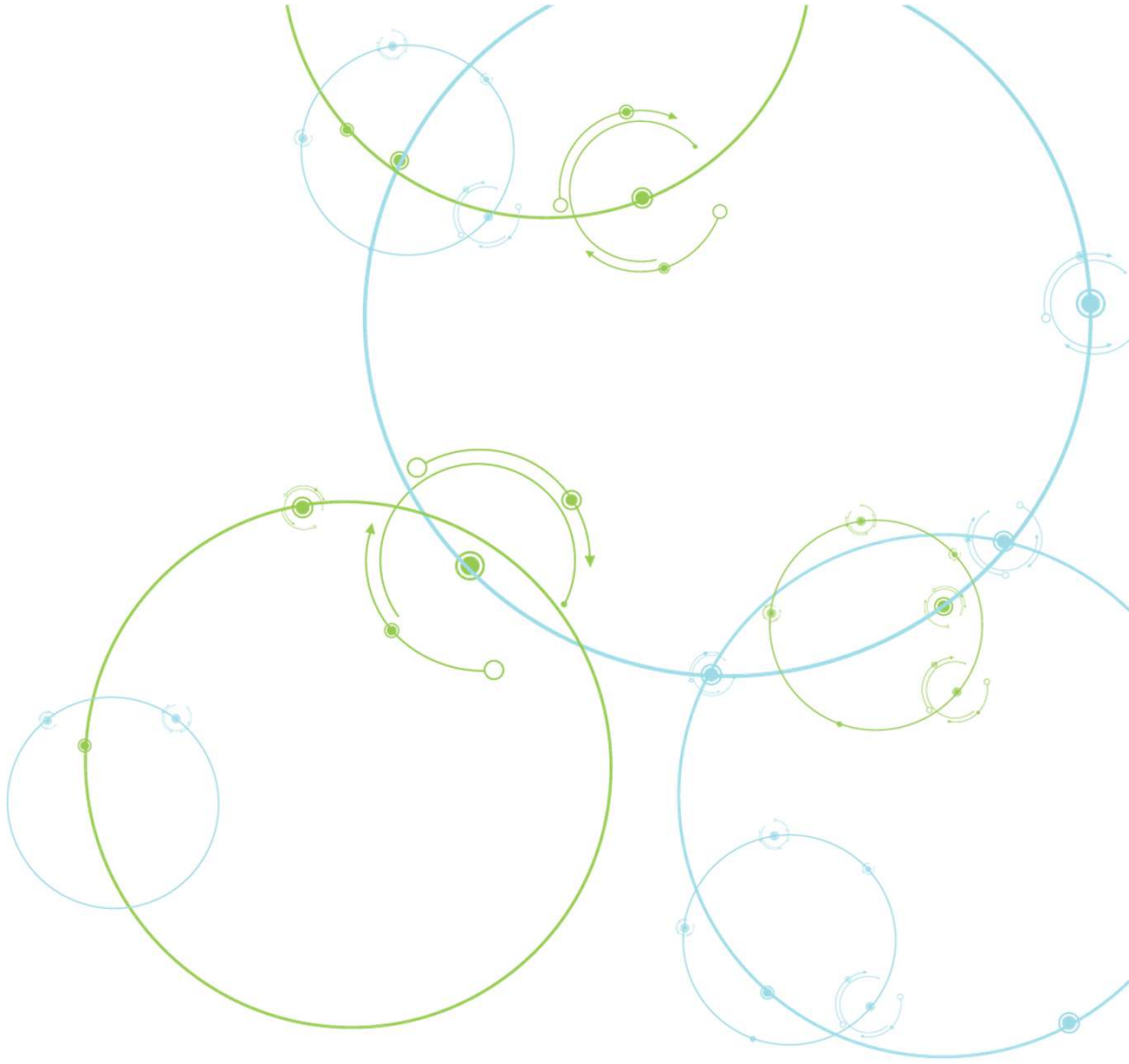
Type of Error	Proportion* human-only	Proportion* AI-assisted
Missed data	6.5%	5.5%
Misallocated data	1.9%	1.8%
Incorrect calculation	0.9%	0.7%
Fabricated data	0.5%	0.8%

*Out of all extracted data items

Median Time on Task

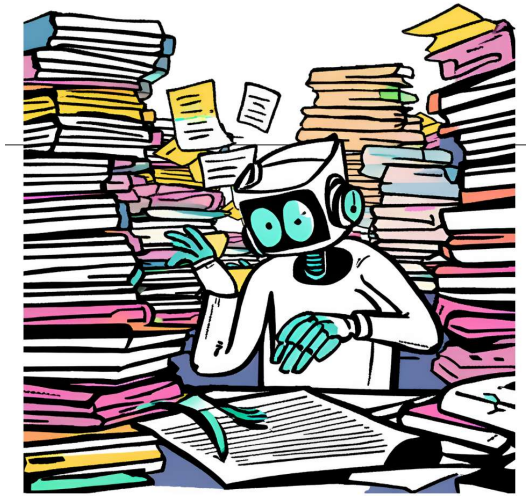


Practical Tips



Practical Tips for AI-assisted Data Extraction

- Several systematic **review platforms** are currently developing AI-assisted data extraction models.
- To date, validation studies evaluating these tools **have not been published**.
- It remains unclear whether these tools will enable users to **customize** underlying prompts.
- **Commercially available LLMs** perform optimally when users have expertise in prompt engineering.



Created by Canva (2025)

Practical Tips for AI-assisted Data Extraction

- Clearly define each data item that needs to be extracted and create a structured list.

  Mean age: The average age of participants in years with standard deviation, reported to one decimal place, overall, and for each treatment group.

  Female participants: The total count and the corresponding percentage, rounded to one decimal place, of female participants in each treatment group.

- Select an LLM with a sufficiently large context window capable of handling uploaded PDFs.

Prompt Engineering-The Art of Effective Inquiry

Created by DALL-E (2025)

- The process of **optimizing input** prompts to effectively guide LLMs in generating **accurate outputs**.
- Provide **clear, direct, and unambiguous** instructions to the LLM.
- **Iteratively test and refine** prompts until the desired extraction output is achieved.

Prompt Engineering-The Art of Effective Inquiry

- **Assign a role**

  You are an expert systematic reviewer specializing in data extraction from academic documents.

- **Assign the task**

  Your task is to carefully read an academic document and extract specific data items.

- **Provide Instructions**

  1. Read the provided academic document thoroughly.
  2. For each item in the data items list:
     a. Search for explicit information in the document.
     b. If found, extract and record the information accurately.
     c. If not found, record it as "not reported".
     d. If possible, calculate any missing numbers or percentages based on available data.

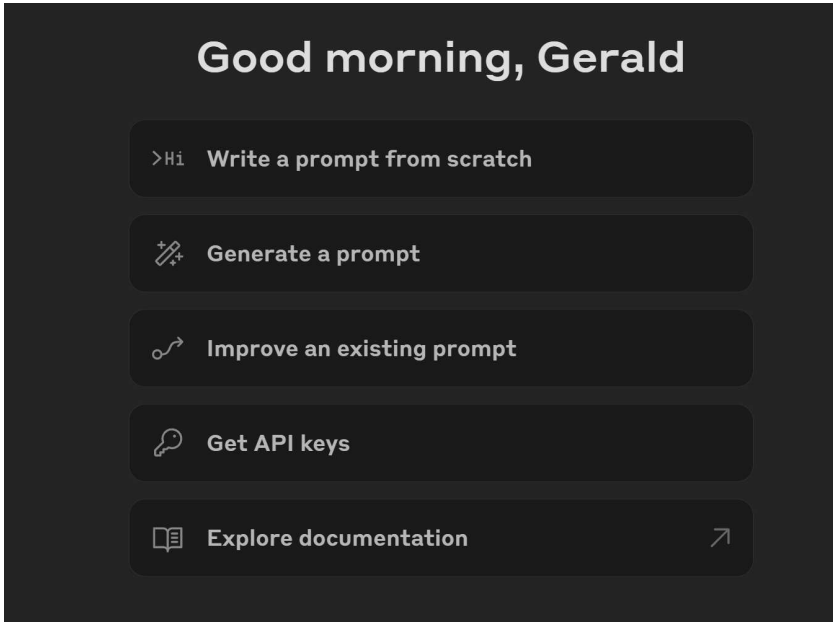
Prompt Engineering-The Art of Effective Inquiry

- **Chain of thought prompting** for complex tasks.
- Provide guidance to the LLM **how to think.** Include **“Think step-by-step“** in the prompt.

If not reported in the study publication, calculate the standard error from the confidence interval of the effect estimate. Think step-by-step before performing the calculation.

1. First determine whether the effect estimate is a mean or a ratio.
2. If it is a ratio, convert the effect estimate and the confidence interval into natural logarithms.
3. Calculate the standard error from the confidence interval.

Support and Courses: https://docs.anthropic.com



- Anthropic prompt builder
- Prompt library
- Interactive courses about prompt engineering

Output

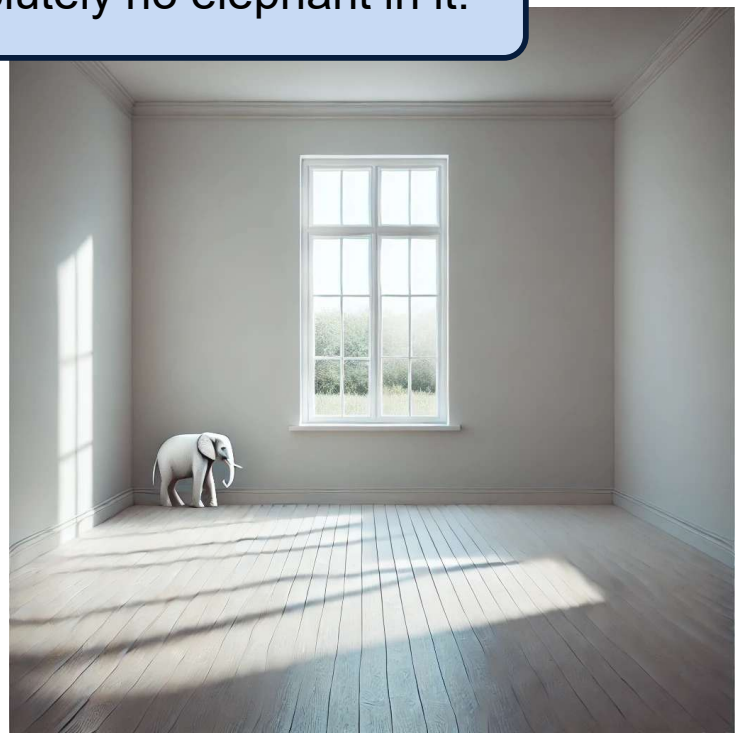
Dual neutralization of both interleukin 17A and interleukin 17F with bimekizumab in patients with psoriasis: Results from BE ABLE 1, a 12-week randomized, double-blinded, placebo-controlled phase 2b trial

Kim A. Papp, MD,[a] Joseph F. Merola, MD,[b] Alice B. Gottlieb, MD, PhD,[c] Christopher E. M. Griffiths, MD,[d] Nancy Cross, MD,[e] Luke Peterson, MS,[e] Christopher Cioffi, PhD,[f] and Andrew Blauvelt, MD[g]
Waterloo, Canada; Boston, Massachusetts; New York, New York; Manchester, United Kingdom; Raleigh, North Carolina; Brussels, Belgium; Portland, Oregon

- **Author's last name:** Papp
- **Year of publication:** 2018
- **Trial registry number:** NCT02905006
- **Study name (acronym):** BE ABLE 1
- **Study design:** Randomized, double-blinded, placebo-controlled, parallel-group, phase 2b trial
- **Study funder:** UCB Pharma
- **Country:** Canada, Czech Republic, Hungary, Japan, Poland, United States
- **Interventions (dosing & route):** Consistent across both documents (bimekizumab SC injections at various doses, every 4 weeks)
- **Study duration:** 12 weeks with additional 20-week safety follow-up
- **Total randomized participants (N=250)** and numbers per treatment group are consistent.
- **Participant characteristics:** Mean age, female participants, population criteria (age, PASI score, BSA involvement), inclusion/exclusion criteria are identical.

Human Oversight is Important

Create a picture of an empty room with absolutely no elephant in it.

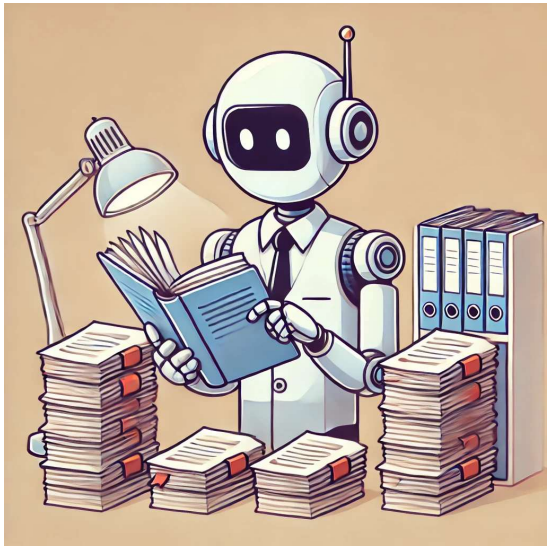


Why is there an elephant in the room?

DALL-E does not inherently favor elephants over other animals; The metaphorical context ("no elephant in the room") can inadvertently prime its visualization, making it more likely to appear in generated imagery.

Conclusions

- The use of an LLM can **improve accuracy and efficiency** in extracting data.
- Human investigators **could be replaced for initial data extraction** if an investigator with experience in prompt engineering for LLMs is available.
- **Prompt engineering** significantly influences the accuracy of data extraction by LLMs.
- **Human oversight** remains essential to verify and validate data extracted by LLMs.



Created by DALL-E (2025)

Acknowledgments

RTI-UNC EPC

- Graham Booth
- Rob Chew
- Karen Crotty
- Rainer Hilscher
- Els Houtsmuller
- Leila Kahwati
- Shannon Kugley
- Meagan Pilar
- Ian Thomas
- Meera Viswanathan

Pacific Northwest EPC

- Steffani Bailey
- Erica Hart
- Rebecca Holmes
- Miranda Pappas
- Carrie Patnode
- Shelley Selph

Minnesota EPC

- Hamdi Abdi
- Sallee Brandt
- Mary Butler
- Amy Claussen
- Toyin Lamina

ECRI

- Jung Min Han
- Benjamin Rouse
- Johnathan Treadwell
- Jesse Wagner

Cochrane Austria

- Andreea Dobrescu
- Claus Nowak
- Barbara Nußbaumer-Streit

Kaiser Permanente EPC

- Sarah Bean
- Erin Coppola
- Carrie Patnode
- Nadia Redmond
- Elizabeth Webber

University of Southern California EPC

- Eric Apaydin
- Shirley Li
- Margaret Maglione
- Bryan Swanton

Thank you

gartlehner@cochrane.at