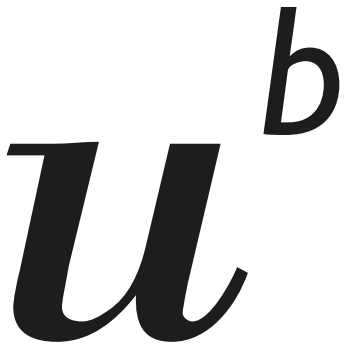


b

UNIVERSITÄT
BERN

A Human-in-the-Loop Scoping Review Screening Pipeline using self-hosted Large Language Models:

An Example in Sport Science

Agenda

- Introduction
  - Research gap
  - Aim
  - Team
- Methods
  - Framework
  - Process
  - Pipeline
- Results
  - Included studies
  - Used resources
  - Human vs AI
- Conclusion

OSF Study protocol: https://osf.io/t8fyh/overview

zenodo Released software: https://zenodo.org/records/20304680

GitHub Software code: https://github.com/KaiGensitz/review-pipeline

BORIS Portal Datasets: https://doi.org/10.48620/97978

Introduction

(just content talk, no code yet)

Introduction

Urbanisation and mobile health

- In 2018, 55% of the world's population resided in urban areas, a figure projected to increase to 68% by 2050 (United Nations, 2018)
- In 2014, there were 1 billion smartphone users; 2026, the number is nearly 4.69 billion and is forecasted to reach 5.83 billion by 2028 (Backlinko Team, 2026)

SOCIETAL CHANGE OF ACTIVITY PATTERNS

Digitalization makes urban everyday life more convenient and let people move less.

- mHealth interventions, often software on smartphones (apps), have small to moderate - short-term - effects in physical activity promotion (Laranjo et al., 2021; Lewis et al., 2024; Mönninghoff et al., 2021; Singh et al., 2024; Stecher et al., 2023)
- But given that: Intervention Impact = Reach X Effectiveness X Receptivity (Abrams et al., 1994; Marcus et al., 2000); mHealth apps have global potential (Aldenaini et al., 2020)

Introduction

Power of artificial intelligence

- mHealth apps harness personal data, having high reach and context-sensitive information (Domin et al., 2021; Küchler et al., 2023) ; but only about actual smartphone users
- Artificial intelligence (AI) can process huge amounts of data (An et al., 2023; Farrahi & Clare, 2024)
- AI algorithms can deliver theory-based stimuli and tasks, enhancing the effectiveness (Alslaity et al., 2023; Canzone et al., 2025; Gabarron et al., 2024; Milne-Ives et al., 2023; Rivera-Romero et al., 2023)
- Level up "Just-in-Time Adaptive Interventions" (Hekler et al., 2018; Nahum-Shani et al., 2018; Wang & Miller, 2020) through AI (e.g., personalizing feature adaptions in real-time; Nahum-Shani & Murphy, 2025)
- But AI is highly energy consuming (Chattaraj & Chimalakonda, 2025; IEA, 2024), though having “green” options (Gutiérrez et al., 2024)

MAIN REVIEW QUESTION

How do smartphone-based AI systems access and promote physical in urban areas?

Research Gap (existing)

Table 1: Previous Reviews regarding mHealth, AI, and PA

Source of evidence (citation)	Year	PA Outcome(s)	mHealth (Smartphone)	AI System(s)	Psychosocial Determinants	Inclusion / Ethics	Sustainability
Domin et al.	2021	✓	✓	✗	✓	✓	✗
An et al.	2023	✓	✓	✓	✗	✓	✗
Brons et al.	2024	✓	✓	✓	✓	✓	✗
Gabarron et al.	2024	✓	✓	✓	✓	✓	✗
Canzone et al.	2025	✓	✓	✓	✗	✓	✗

Note: AI = Artificial Intelligence, mHealth = mobile Health, PA = Physical Activity

Aim

The aim is to summarize how **smartphone applications** based on **artificial intelligence** assess and promote **physical activity** in **urban areas**.

The review will explore to what extent the applications integrate **psychosocial factors**, consider **inclusion of diverse groups** and **ethics** as well as **sustainability** in the development and implementation process.

Research Team

- **Kai M. Gensitz**, Department of Health Science, Institute for Sport Science, University of Bern, Switzerland, E-Mail: kai.gensitz@unibe.ch
- **Shawan Mohammed**, Chair for Distributed Signal Processing, RWTH Aachen University, Germany, E-Mail: mohammed@ice.rwth-aachen.de
- **Daniela E. Ströckl**, Medical Informatics, University of Applied Science Kärnten, E-Mail: D.Stroeckl@fh-kaernten.at
- **Marc Augustin**, Social Medizin, EvH Bochum, Germany, E-Mail: marc.augustin@evh-bochum.de
- **Claudio R. Nigg**, Department of Health Science, Institute for Sport Science, University of Bern, Switzerland, E-Mail: clauido.nigg@unibe.ch
- **Ciara McCormack**, Department of Sport Science and Nutrition, National University of Ireland Maynooth, E-Mail: ciara.a.mccormack@mu.ie

Review process

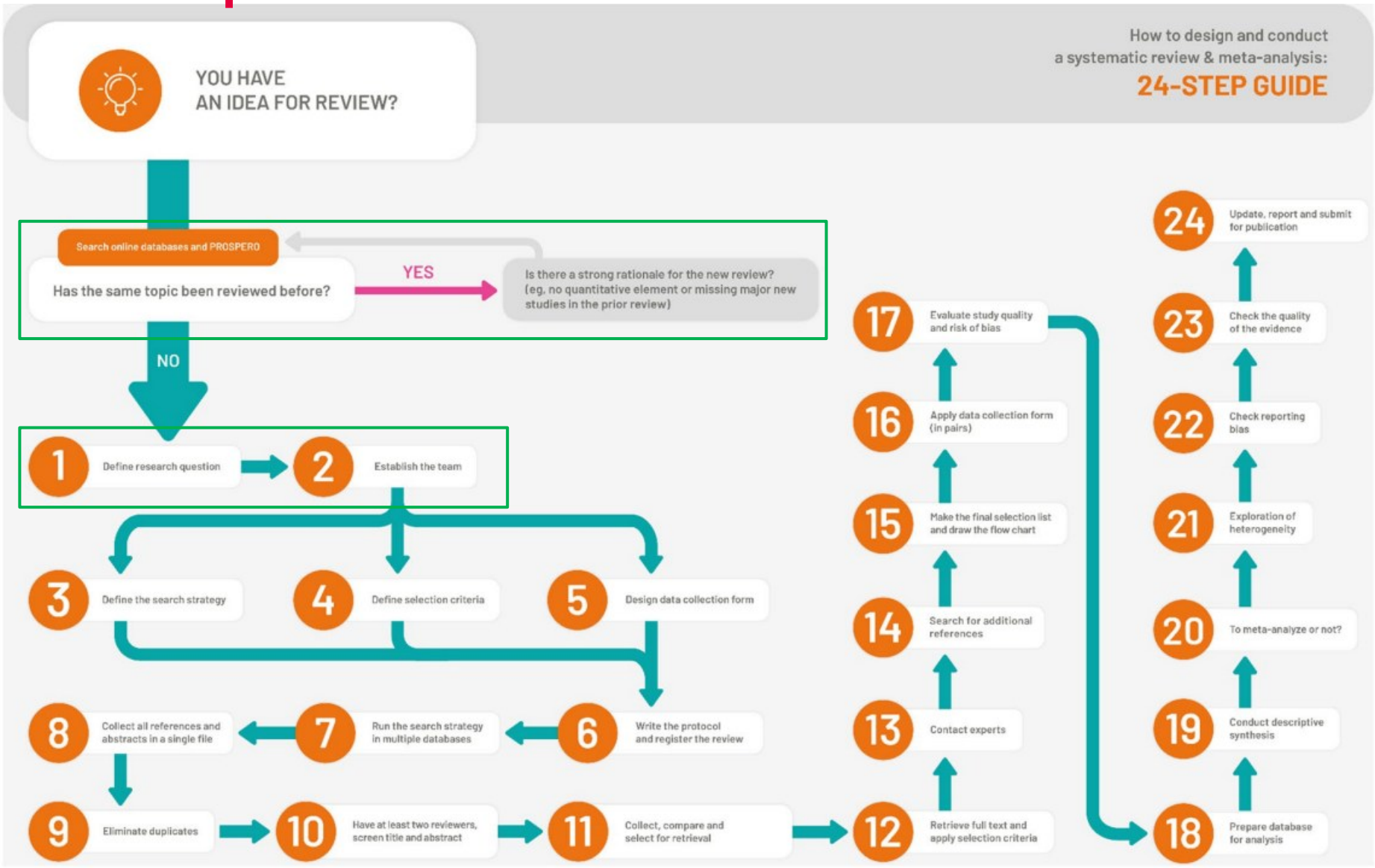


Figure 1: 24-Step Guide for Systematic Reviews (Muka et al., 2020)

Research Gap (closed)

Table 2: Previous Reviews and Upcoming Review regarding mHealth, AI, and PA

Source of evidence (citation)	Year	PA Outcome(s)	mHealth (Smartphone)	AI System(s)	Psychosocial Determinants	Inclusion / Ethics	Sustainability
Domin et al.	2021	✓	✓	✗	✓	✓	✗
An et al.	2023	✓	✓	✓	✗	✓	✗
Brons et al.	2024	✓	✓	✓	✓	✓	✗
Gabarron et al.	2024	✓	✓	✓	✓	✓	✗
Canzone et al.	2025	✓	✓	✓	✗	✓	✗
Gensitz et al.	2026†	✓	✓	✓	✓	✓	✓

Note: AI = Artificial Intelligence, mHealth = mobile Health, PA = Physical Activity; † under review

Methods

(equally content and code talk)

Determining review framework

- Scoping Review following PRISMA extension (PRISMA-ScR; Tricco et al., 2018): overview about research field (von Elm et al., 2019) to identify existing knowledge (Peters et al., 2022)
- Study protocol for preregistration: Eligibility criteria with **PCC** scheme (Peters et al., 2022)
  - **P**opulation: All adults (18+)
  - **C**oncepts (article should include outcome, intervention, phenomenon, and have implementation information):
    a) **Outcome**: Physical Activity Behavior
    b) **Intervention**: Mobile Health via Smartphone
    c) **Phenomenon**: Artificial Intelligence Systems
    d) **Implementation**: System Idea and Development Process
       1) Psychosocial Determinants
       2) Target Group Inclusion and Ethics
       3) Sustainability Considerations
  - **C**ontext: Urban areas globally
  - Types of evidence sources: (protocols of) primary studies

Review process

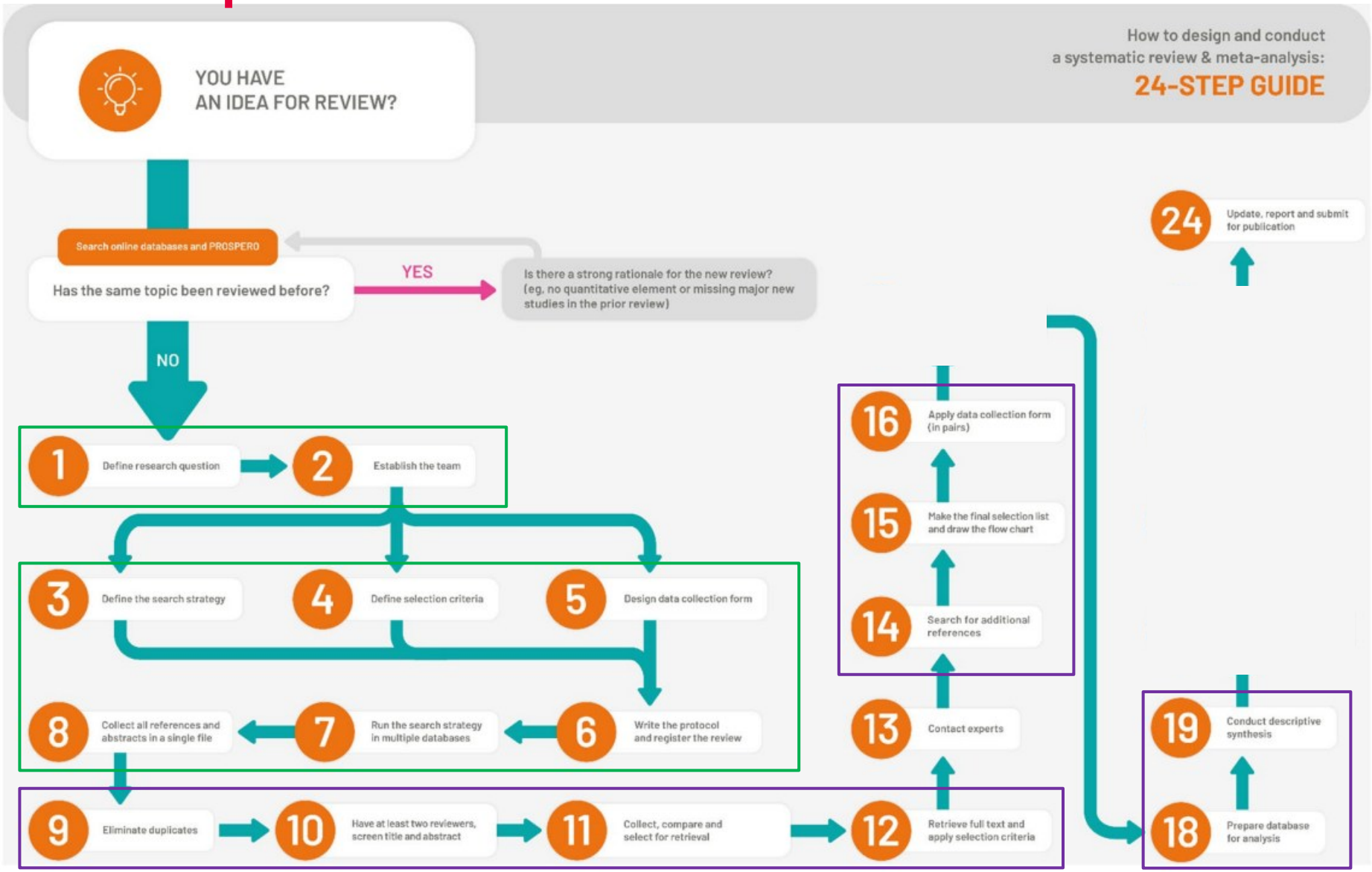


Figure 1: 24-Step Guide for Systematic Reviews (Muka et al., 2020)

Used software for review process

- **Reference workflow**: citation management in Zotero (Stillman & Cheslack-Postava, 2026)
- **Human screening**: calibration, conflict resolution, and reviewer accountability in Covidence (Veritas Health Innovation, 2026)
- **Pipeline coding**: vibe-coded with ChatGPT Codex VS Code extension 5.5; human-on-the-loop review, correction, and specification (OpenAI, 2026)
- **Pipeline execution:** self-hosted inference
  - secure API calls through GPUStack on University of Bern servers (GPUStack.ai, 2022)
  - Models: gpt-oss-120b (OpenAI, 2025) for screening/extraction decisions; qwen3-embedding-0.6b (Qwen, 2025) for chunking and retrieval
- **Citation searching**: seed references processed with citationchaser (Haddaway et al., 2021)

Different ways of using AI in the project

SCIENCE

Review execution

Human-in-the-loop

- Eligibility criteria are human-defined
- Calibration/test sample are human-screened first
- Conflicts and threshold are human-adjudicated
- Pipeline perform repeatable work
- Outputs are auditable

humans decide; pipeline works

SOFTWARE

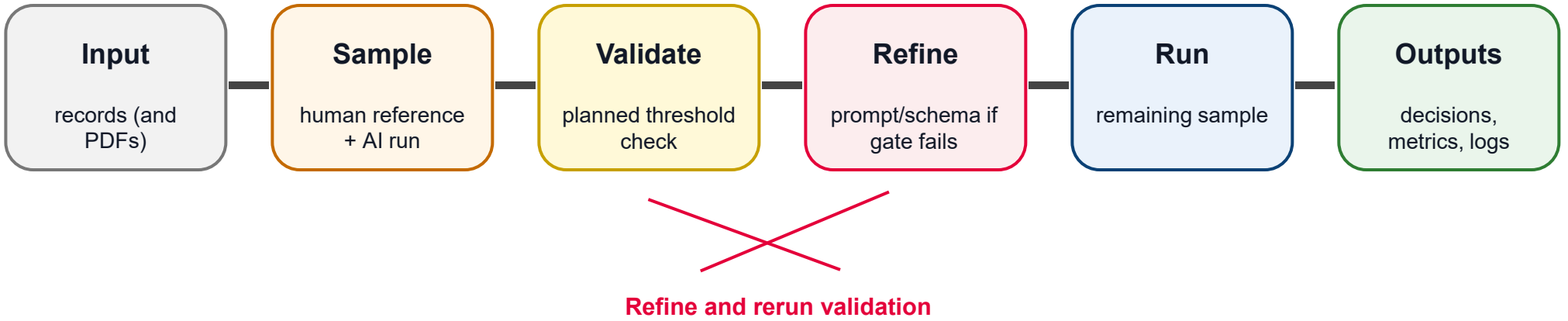
Code development

Human-on-the-loop

- Codex/ChatGPT produced implementation drafts
- Human reviewed direction and errors
- Human intervened for failures or design choices
- Final accountability stays with the human experts

AI worked and often chose; humans supervised and corrected

Workflow: AI-assisted review pipeline



Title/Abstract	10% calibration + 5% test set	**Sensitivity >=95%; PABAK >=80%**
Full-Text	test sample with PDF evidence	**Sensitivity >=95%; PABAK >=80%**
Data Extraction	4 papers, many fields	**Concordance >=80%; Accuracy >=90%**

PABAK = prevalence-adjusted and bias-adjusted kappa (Byrt et al., 1993)

Results

(little content and much code talk)

Flow diagramm of included studies

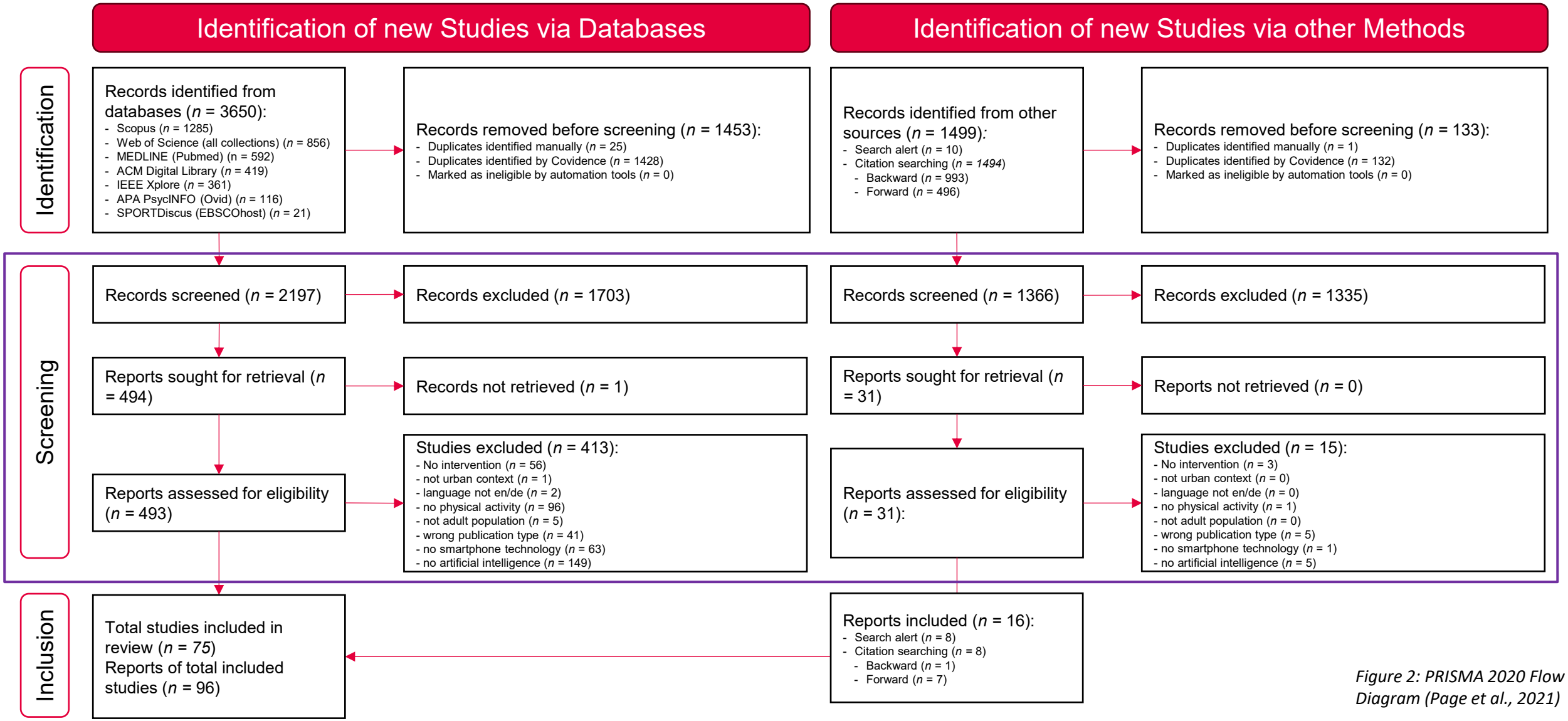


Figure 2: PRISMA 2020 Flow Diagram (Page et al., 2021)

Decision Matrix: Title/Abstract

Title/Abstract calibration set (n=223)

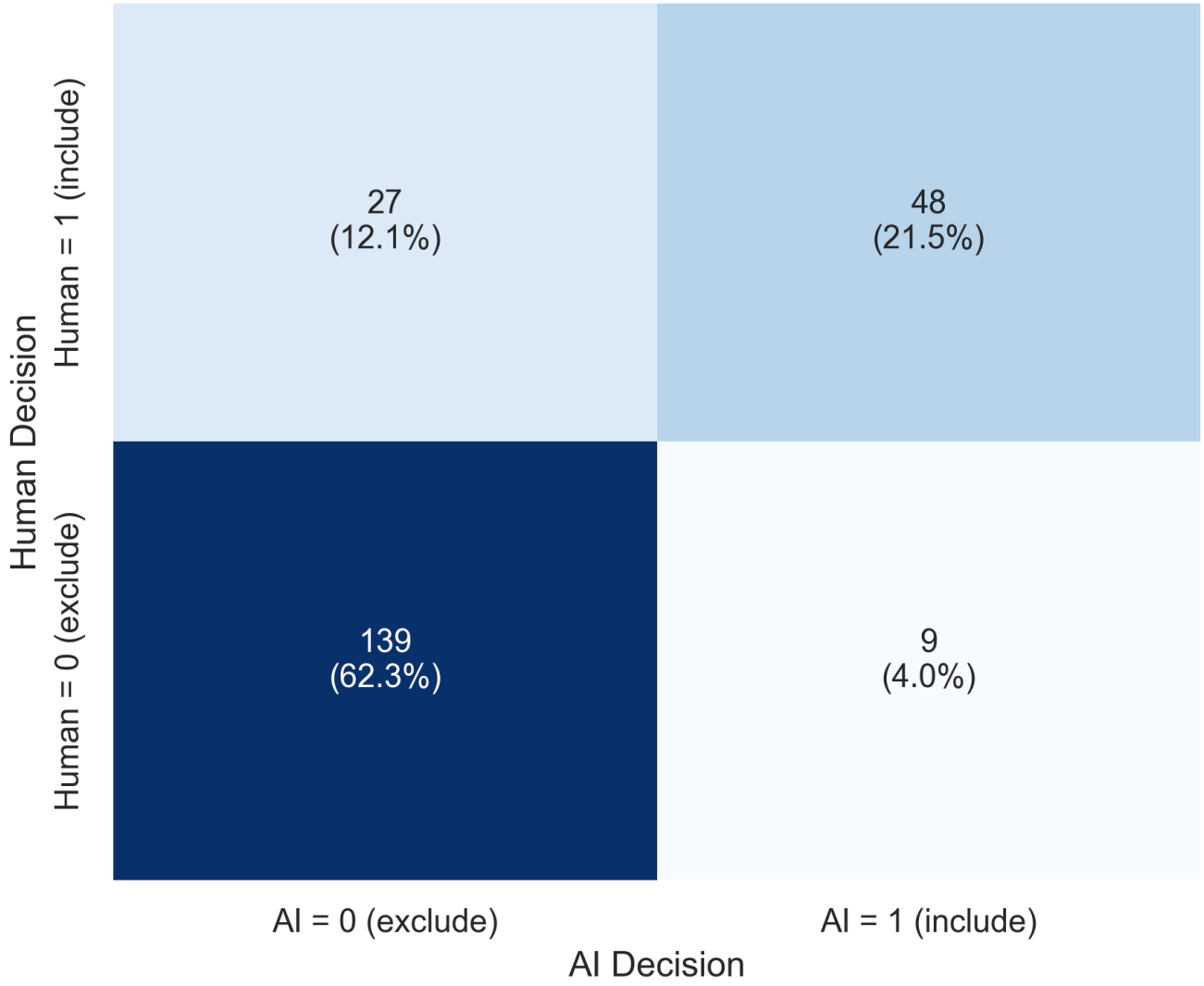


Figure 3.1: Comparison of false positive/negative hits from AI (Title/Abstract)

Decision Matrix: Title/Abstract

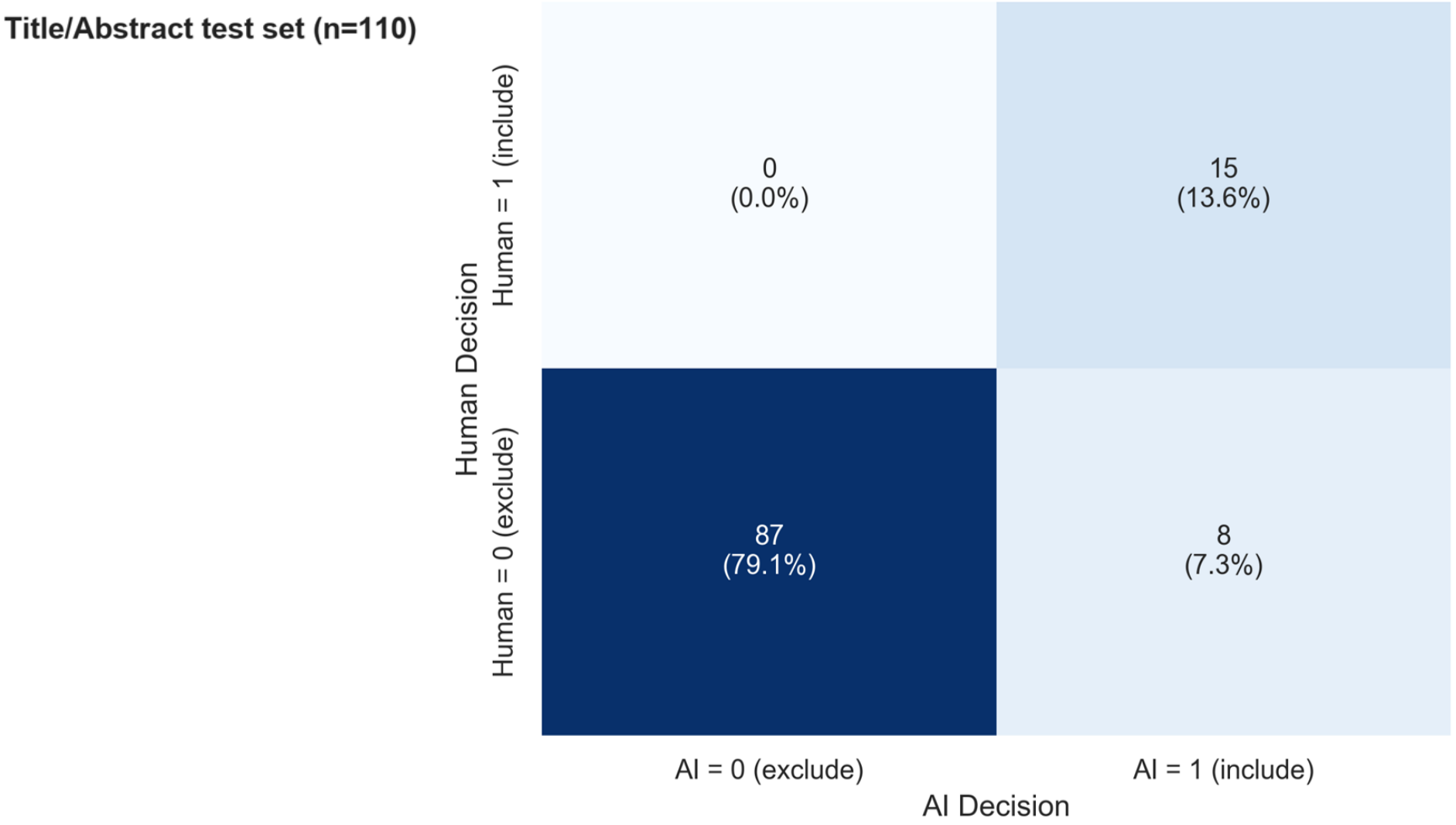


Figure 3.2: Comparison of false positive/negative hits from AI (Title/Abstract)

Decision Matrix: Full-Text

Full-Text test set (n=50)

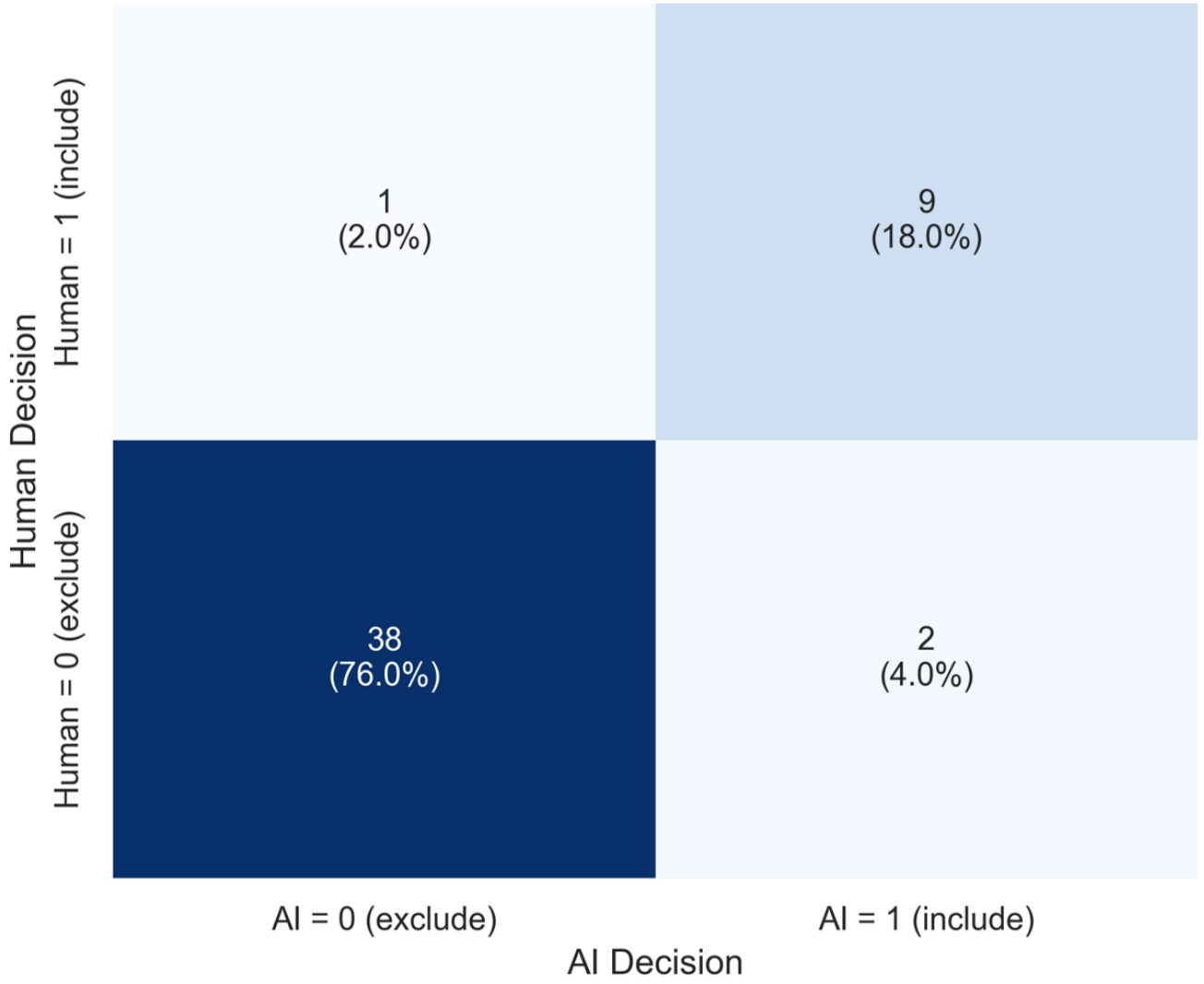


Figure 3.3: Comparison of false positive/negative hits from AI (Full-Text)

Flow diagramm of included studies

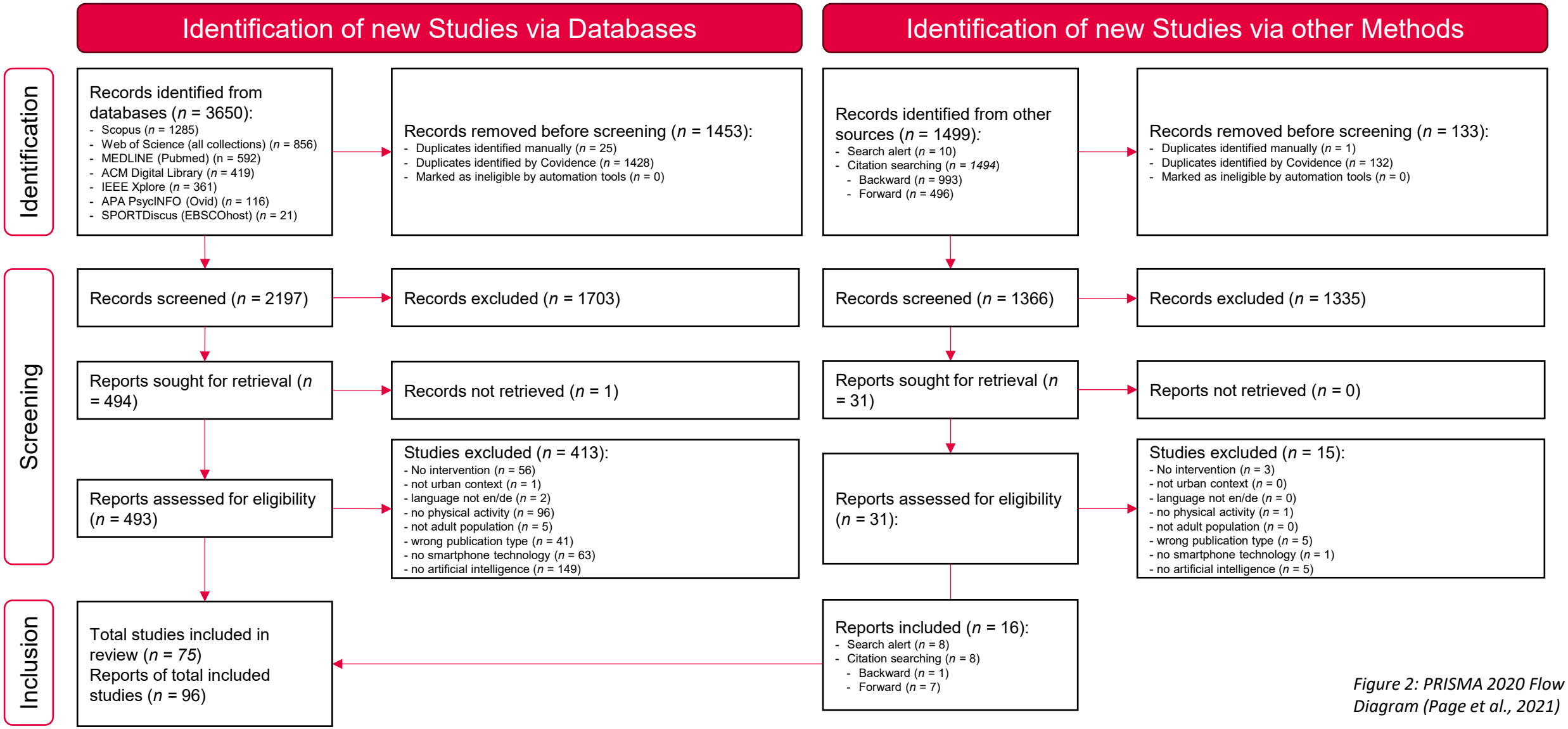
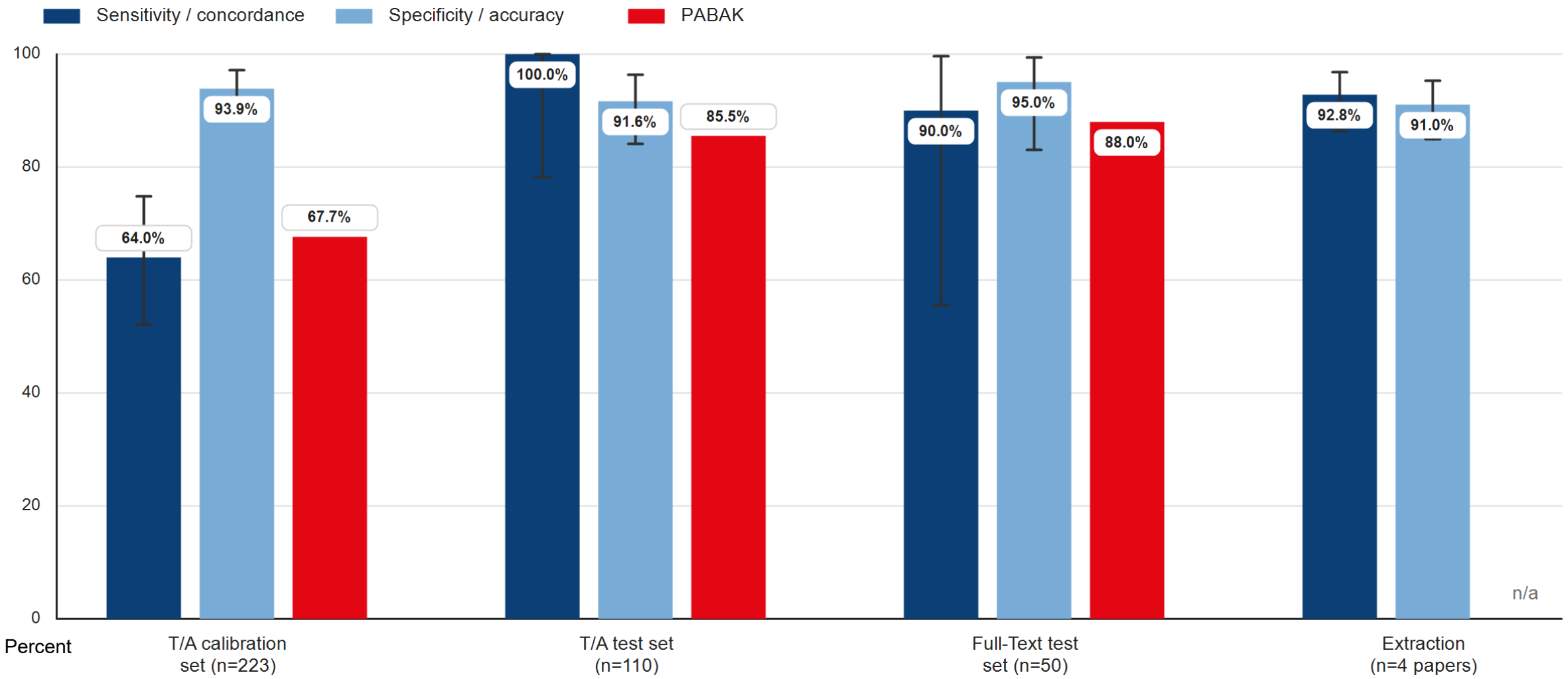


Figure 2: PRISMA 2020 Flow Diagram (Page et al., 2021)

AI validation across all review stages



Error bars show 95% CIs; T/A = Title/Abstract, PABAK = prevalence-adjusted and bias-adjusted kappa

Figure 4: Validation thresholds comparison (Human Consensus vs AI)

What we found – content wise

Smartphone AI for Physical Activity in Urban Areas

- **96 reports / 75 unique approaches**: 35/75 positive physical activity, 10/75 mixed/null/declining effects, others had no outcome reporting
- **Context**: setting rarely (19/75) specified (15 urban, 4 urban/rural)
- **AI systems**: machine learning (28), reinforcement learning (18), rule-based hybrid (17), deep learning (16), generative AI (11), recommender systems (9)
- **Psychosocial theory**: Behavior change techniques were commonly reported (64) but for 41/75 study approaches no exact theory to AI linkage was possible
- **Inclusion/ethics**: 69/75 mentioned consent/ethics review, while only 9 addressed equity/accessibility/underserved groups and only 5 reported algorithmic fairness, explainability, interpretability, or bias safeguards
- **Sustainability**: 35/75 mentioned scalability/cost, but only 15 reported ecological impacts, mostly battery or low-power sensing

Token usage: Title/Abstract stage

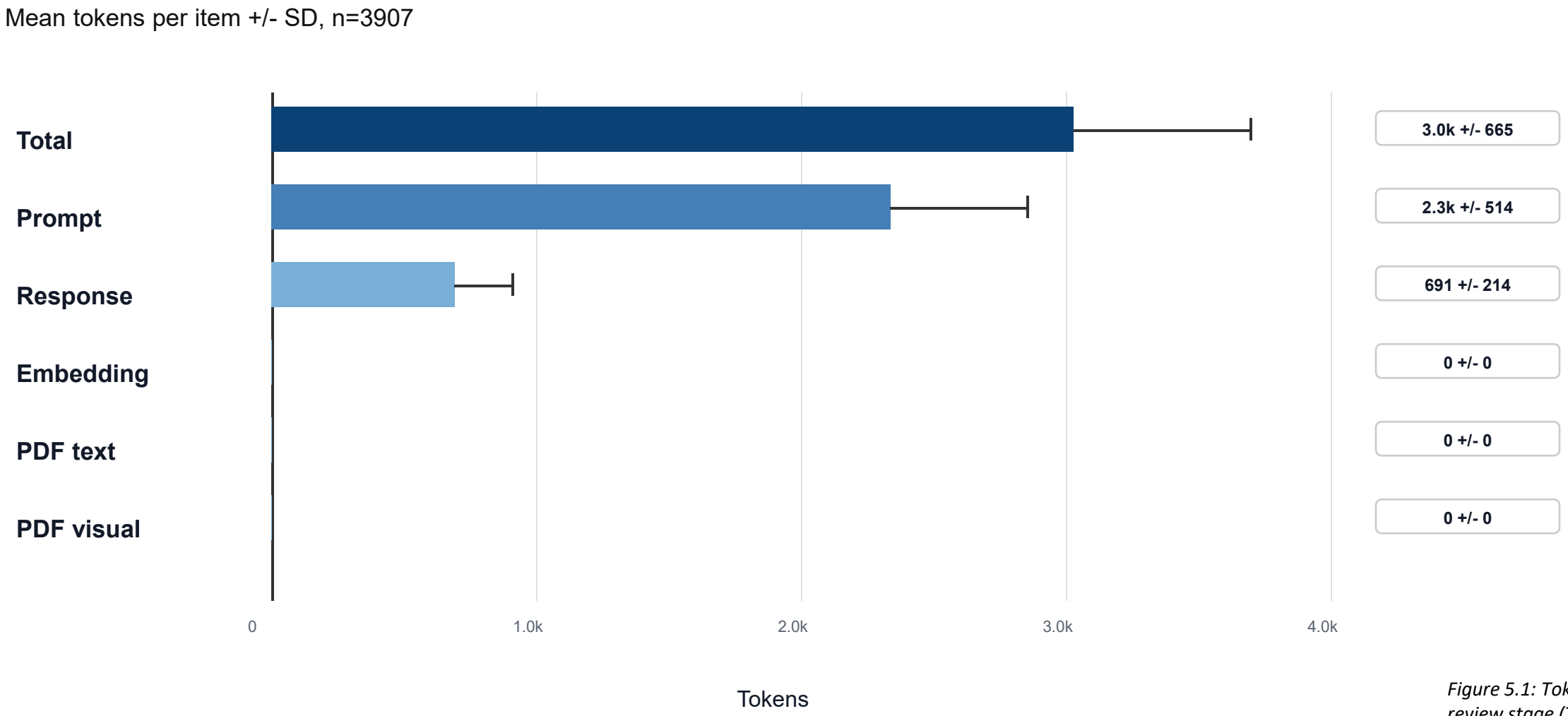


Figure 5.1: Token usage per review stage (Title/Abstract)

Token usage: Full-Text stage

Mean tokens per item +/- SD, n=516. PDF text, visual, and embedding tokens appear only in this stage.

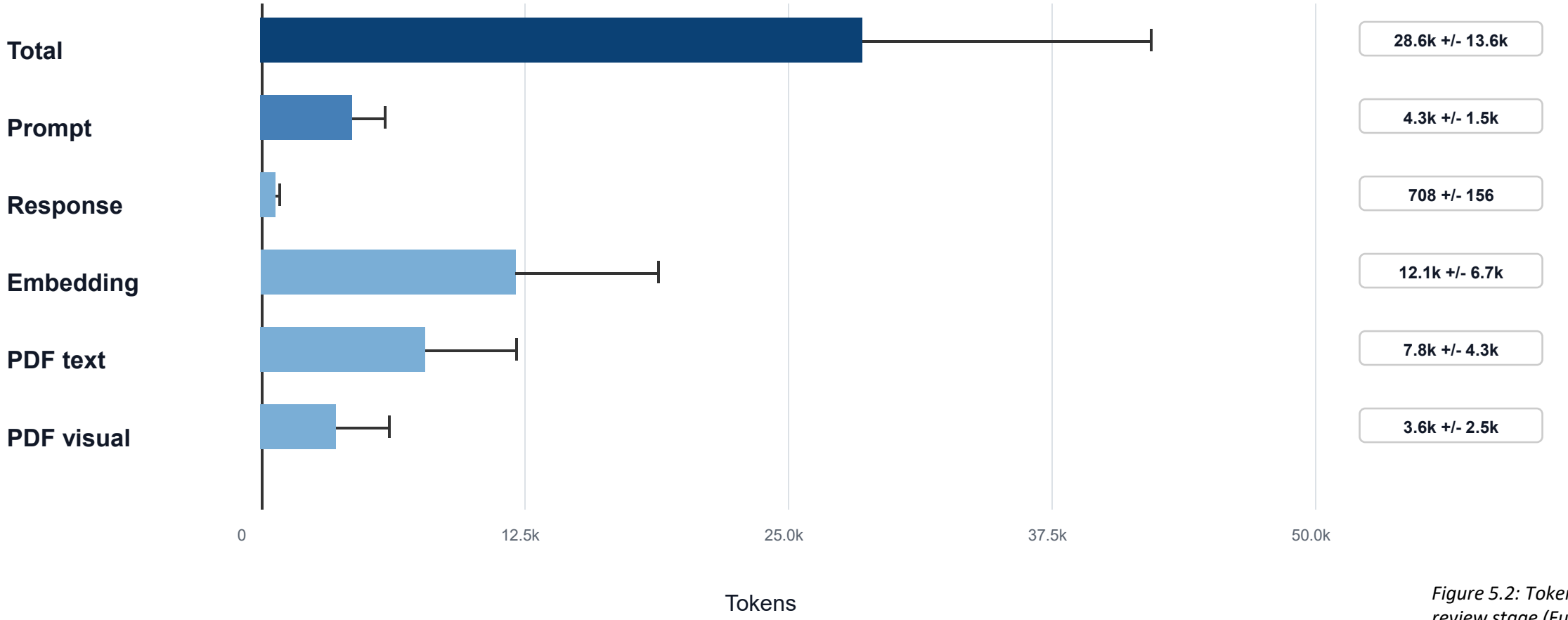


Figure 5.2: Token usage per review stage (Full-Text)

Token usage: Data Extraction stage

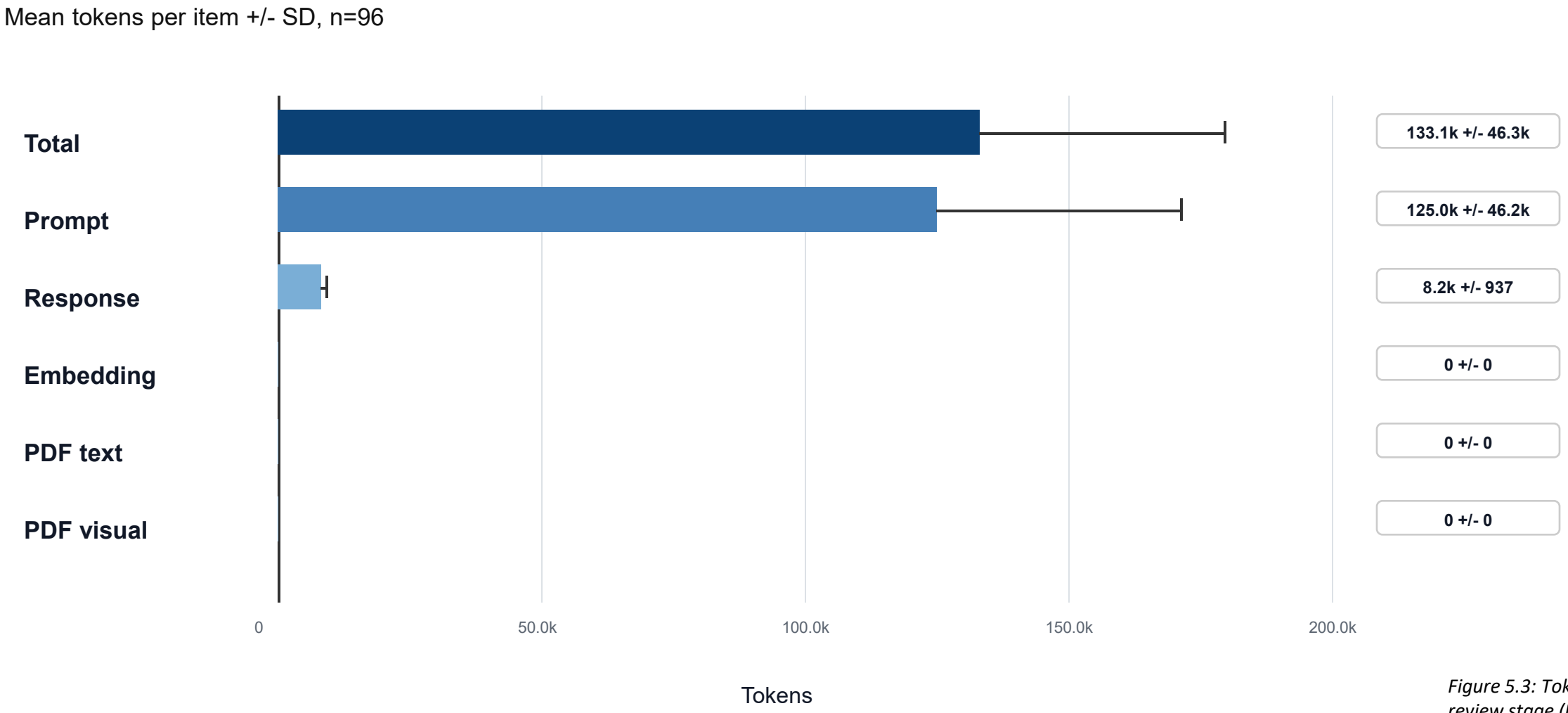
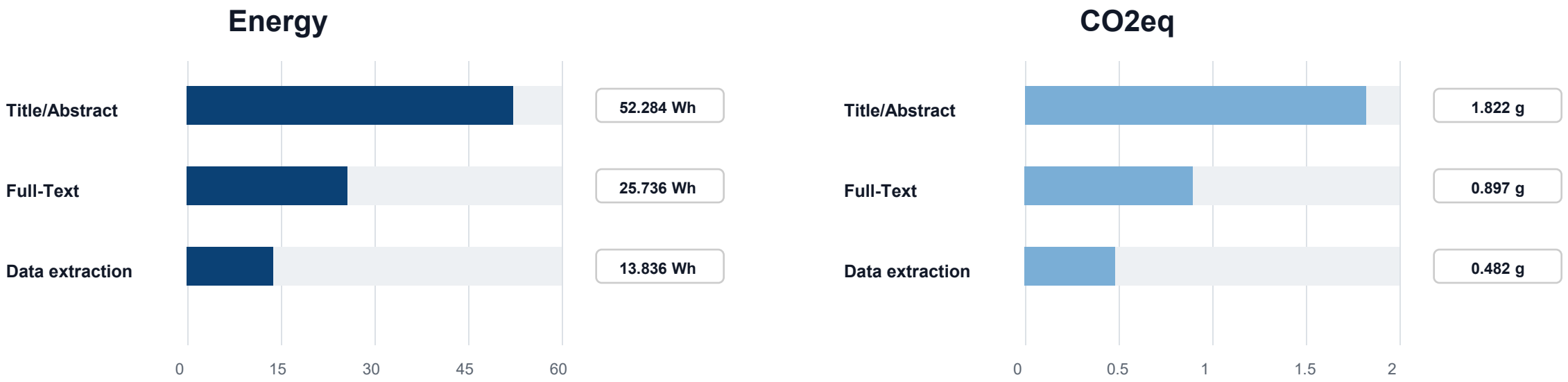


Figure 5.3: Token usage per review stage (Data Extraction,

Local operational resources logged



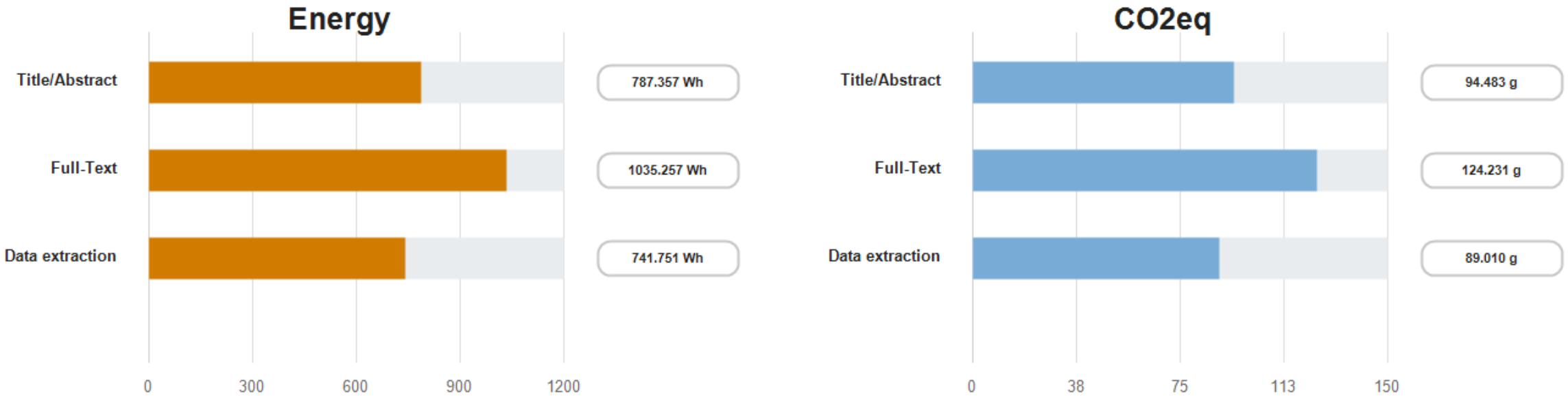
Total: 0.092 kWh | 0.003 kg CO2eq

Figure 6.1: Resource usage from pipeline (Local Machine)

- CodeCarbon further logged water consumed = 0.0 and WUE = 0.0
- Human-only comparator: about 21.5 kWh metabolic energy for typing-like deskwork

Excludes software development and testing, prompt engineering, setup work, embodied hardware, institutional overhead, and any unlogged citation-searching savings.

Estimated resources used with UBELIX



Total: 2.564 kWh | 0.308 kg CO2eq

Figure 6.2: Resource usage from pipeline (Server)

- UBELIX estimate uses H100 TDP = 350 W, PUE = 1.580, grid intensity = 120 g CO2eq/kWh
- IT energy before PUE = 1.623 kWh; estimates gpt-120b-oss and qwen3-embedding-0.6b via GPUStack service

Excludes software development and testing, prompt engineering, setup work, embodied hardware, institutional overhead, and any unlogged citation-searching savings.

Estimated resources used in total

UBELIX facility values are only estimates and not directly measured during the pipeline process

Excludes software development and testing, prompt engineering, setup work, embodied hardware, institutional overhead, and any unlogged citation-searching savings.

Figure 6.3: Resource usage from pipeline (Total)

Logged time: humans vs AI

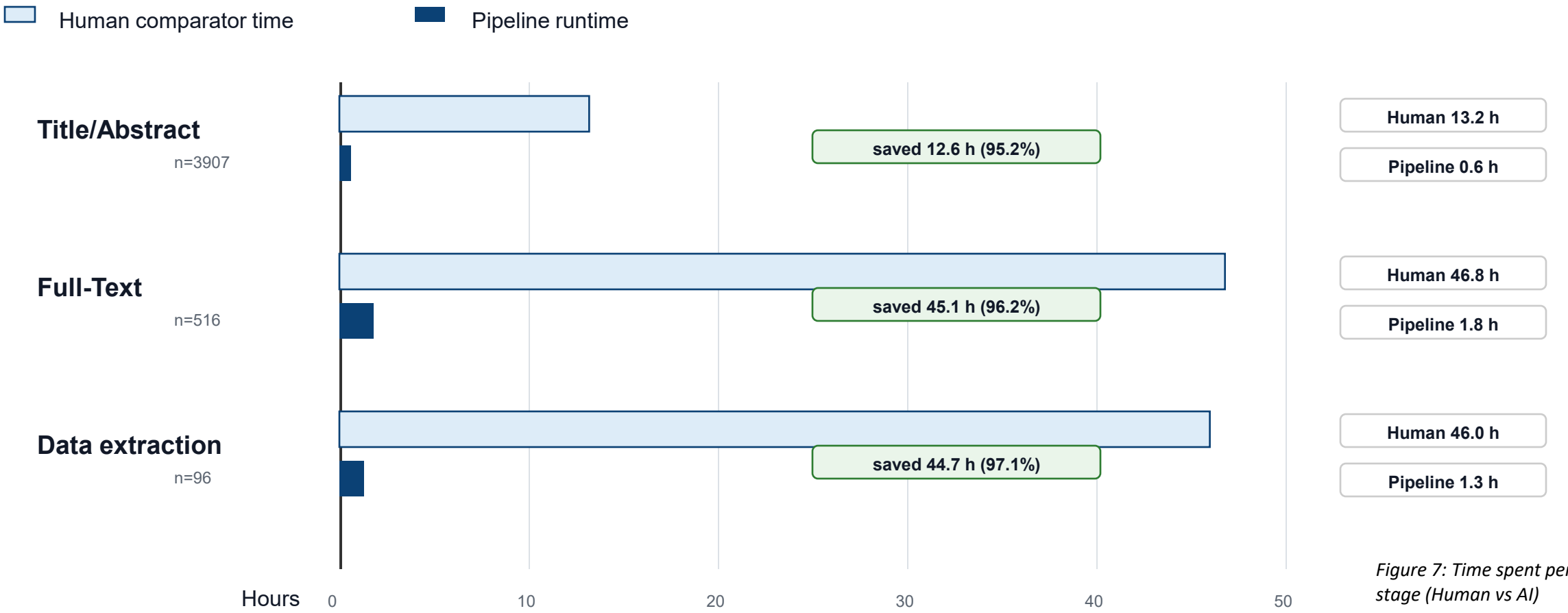


Figure 7: Time spent per review stage (Human vs AI)

Total logged reviewer time saved: 102.3 h | Transfer/setup: 13.4 h | Pipeline-only potential: 115.7 h

Conclusion

(equally content and code talk)

Findings – content wise

Smartphone AI for physical activity in urban areas

- **Outcomes and context**: effects for physical activity are promising but uneven and need durability, dropout, and subgroup testing; setting is often underspecified
- **Theory and AI logic**: behavior change techniques are common but theory linkage is weak; document inputs, policies, confidence, and fallback logic
- **Equity and inclusitivity**: Ethics approval reporting is not enough; pre-register analyses, accessibility needs, bias checks, and oversight procedures
- **Sustainability**: Resource use is rarely measured; report battery drain, data transfer, compute location, kWh, and CO2eq would be advised

Limitations – content wise

Smartphone AI for physical activity in urban areas

- **Heterogeneous evidence:** 96 reports mixed RCTs, protocols, feasibility studies, technical evaluations, and secondary analyses; findings map patterns but we cannot compare effectiveness
- **Study clustering uncertainty:** 75 unique approaches were reconstructed from reports, but overlap may remain when related protocols, algorithms, and outcome papers were unclear
- **Incomplete reporting:** setting, subgroup characteristics, algorithm transparency, equity safeguards, and sustainability metrics were often missing or weakly specified

FUTURE RESEARCH QUESTION

For whom, and with which theoretical mechanisms as well as equity and sustainability considerations can AI-based mHealth interventions meaningfully support physical activity?

Findings – code wise

AI-assisted review workflow

- **Feasible workflow:** Self-hosted AI models can support title/abstract screening, full-text screening, and data extraction within auditable scoping review pipeline
- **Human accountability:** currenlty only review support, not replacement; humans defined criteria, validate samples, adjudicate conflicts, and approve stage gates
- **Validation:** confusion matrices, field-level extraction checks, and resource logs make AI assistance inspectable and are in place
- **Next steps:** repeat the process across topics and institutions to learn when self-hosted inference is reliable and where it fails

Limitations – code wise

AI-assisted review workflow

- **Sampled validation only:** good agreement was shown in validation samples, but false negatives and extraction errors remain possible in the full corpus
- **Model- and workflow-dependent:** results are dependent on parsing quality, prompt/schema choices, English/German language policy, and the specific self-hosted large language and embedding models with limited size used

FUTURE OPPORTUNITY

Team-up with interested researchers and developers to make the AI-assisted review pipeline a user-friendly software for the whole scientific community.

Contact

Kai Gensitz, M.Sc.

Doctoral Candidate

kai.gensitz@unibe.ch

+41 31 684 87 12

OSF Study protocol: https://osf.io/t8fyh/overview

zenodo Released software: https://zenodo.org/records/20304680

GitHub Software code: https://github.com/KaiGensitz/review-pipeline

BORIS Portal Datasets: https://doi.org/10.48620/97978

References

Aldenaini, N., Alqahtani, F., Orji, R., & Sampalli, S. (2020). Trends in Persuasive Technologies for Physical Activity and Sedentary Behavior: A Systematic Review. *Frontiers in Artificial Intelligence*, *3*, 7. https://doi.org/10.3389/frai.2020.00007

Alslaity, A., Chan, G., & Orji, R. (2023). A panoramic view of personalization based on individual differences in persuasive and behavior change interventions. *Frontiers in Artificial Intelligence*, *6*, 1125191. https://doi.org/10.3389/frai.2023.1125191

An, R., Shen, J., Wang, J., & Yang, Y. (2023). A scoping review of methodologies for applying artificial intelligence to physical activity interventions. *Journal of Sport and Health Science*. https://doi.org/10.1016/j.jshs.2023.09.010

Backlinko Team. (2026, January 16). Smartphone Usage Statistics [Blacklinko]. https://backlinko.com/smartphone-usage-statistics, accessed: 2026, April 21

Barisch-Fritz, B., Nigg, C. R., Barisch, M., & Woll, A. (2022). App development in a sports science setting: A systematic review and lessons learned from an exemplary setting to generate recommendations for the app development process. *Frontiers in Sports and Active Living*, *4*, 1012239. https://doi.org/10.3389/fspor.2022.1012239

Bosch. (2018, January 30). Die Geschichte der Künstlichen Intelligenz: Von Turing bis Watson: Die Entwicklung der denkenden Systeme. *Künstliche Intelligenz*. https://www.bosch.com/de/stories/geschichte-der-kuenstlichen-intelligenz/

Bramer, W. M., Rethlefsen, M. L., Kleijnen, J., & Franco, O. H. (2017). Optimal database combinations for literature searches in systematic reviews: A prospective exploratory study. *Systematic Reviews*, *6*(1), 245. https://doi.org/10.1186/s13643-017-0644-y

Brons, A., Wang, S., Visser, B., Kröse, B., Bakkes, S., & Veltkamp, R. (2024). Machine Learning Methods to Personalize Persuasive Strategies in mHealth Interventions That Promote Physical Activity: Scoping Review and Categorization Overview. *Journal of Medical Internet Research*, *26*(1), e47774. https://doi.org/10.2196/47774

Canzone, A., Belmonte, G., Patti, A., Vicari, D. S. S., Rapisarda, F., Giustino, V., Drid, P., & Bianco, A. (2025). The multiple uses of artificial intelligence in exercise programs: A narrative review. *Frontiers in Public Health*, *13*, 1510801. https://doi.org/10.3389/fpubh.2025.1510801

Chattaraj, R., & Chimalakonda, S. (2025). NLP Libraries, Energy Consumption and Runtime: An Empirical Study. *Proc. ACM Softw. Eng.*, *2*(FSE), FSE126:2850-FSE126:2873. https://doi.org/10.1145/3729396

Domin, A., Spruijt-Metz, D., Theisen, D., Ouzzahra, Y., & Vögele, C. (2021). Smartphone-Based Interventions for Physical Activity Promotion: Scoping Review of the Evidence Over the Last 10 Years. *JMIR mHealth and uHealth*, *9*(7), 24308. https://doi.org/10.2196/24308

References

Farrahi, V., & Clare, P. (2024). Artificial Intelligence and Machine Learning-Powerful Yet Underutilized Tools and Algorithms in Physical Activity and Sedentary Behavior Research. *Journal of Physical Activity & Health*, 1–3. https://doi.org/10.1123/jpah.2024-0021

Gabarron, E., Larbi, D., Rivera-Romero, O., & Denecke, K. (2024). Human Factors in AI-Driven Digital Solutions for Increasing Physical Activity: Scoping Review. *JMIR Human Factors*, *11*(1), e55964. https://doi.org/10.2196/55964

GPUStack.ai. (2022). GPUStack (Version 2.1.1) [Open-source GPU cluster manager]. University of Bern. https://gpustack.unibe.ch/

Gutiérrez, M., Moraga, M. Á., García, F., & Calero, C. (2024). Green IN Artificial Intelligence from a Software Perspective: State-of-the-Art and Green Decalogue. *ACM Comput. Surv.*, *57*(3), 64:1-64:30. https://doi.org/10.1145/3698111

Haddaway, N. R., Grainger, M. J., & Gray, C. T. (2021). citationchaser: An R package and Shiny app for forward and backward citations chasing in academic searching (Version 0.0.3) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.4543513

Hekler, E. B., Rivera, D. E., Martin, C. A., Phatak, S. S., Freigoun, M. T., Korinek, E., Klasnja, P., Adams, M. A., & Buman, M. P. (2018). Tutorial for Using Control Systems Engineering to Optimize Adaptive Mobile Health Interventions. *Journal of Medical Internet Research*, *20*(6), 214. https://doi.org/10.2196/jmir.8622

Hirt, J., Nordhausen, T., Fuerst, T., Ewald, H., & Appenzeller-Herzog, C. (2024). Guidance on terminology, application, and reporting of citation searching: The TARCiS statement. *BMJ*, *385*, e078384. https://doi.org/10.1136/bmj-2023-078384

IEA. (2024). *World Energy Outlook 2024 – Analysis*. https://www.iea.org/reports/world-energy-outlook-2024, accessed at 23.06.2025

Küchler, A.-M., Ebert, D. D., & Baumeister, H. (2023). Körperliche Aktivität. In D. D. Ebert & H. Baumeister (Eds.), *Digitale Gesundheitsinterventionen: Anwendungen in therapie und* (pp. 207–225). SPRINGER-VERLAG BERLIN AN. https://doi.org/10.1007/978-3-662-65816-1_12

Laranjo, L., Ding, D., Heleno, B., Kocaballi, B., Quiroz, J. C., Tong, H. L., Chahwan, B., Neves, A. L., Gabarron, E., Dao, K. P., Rodrigues, D., Neves, G. C., Antunes, M. L., Coiera, E., & Bates, D. W. (2021). Do smartphone applications and activity trackers increase physical activity in adults? Systematic review, meta-analysis and metaregression. *British Journal of Sports Medicine*, *55*(8), 422–432. https://doi.org/10.1136/ bjsports-2020-102892