جامعة ابن طفيل
ⵜⴰⵙⴷⴰⵡⵉⵜ ⵉⴱⵏ ⵟⵓⴼⴰⵢⵍ
Ibn Tofaïl University

Faculté des Sciences

Université Ibn Tofail
Faculté des Sciences, Kénitra
Mémoire de Projet de Fin d'Études
Master Intelligence Artificielle et Réalité Virtuelle

Systematic Reviews and Artificial Intelligence: Mechanisms for Success and Failure

---

Établissement d'accueil : Open Development & Education- (UK)

Elaboré par : Mr. : TOUIYATE Adnane

Encadré par : Mr. : BOUJIHA Tarik (ENSA-UIT)

Mr. : MANSOUR Hassan (OpenDevEd)

Soutenu le 20 septembre 2024, devant le jury composé de :

- Pr. Rochdi Messoussi (FSK Université Ibn Tofaïl)
- Pr. Rajae Touahni (FSK Université Ibn Tofaïl)
- Pr. Tarik Boujiha (ENSA-UIT)
- Pr. Anas Nouri (FSK Université Ibn Tofaïl)
- Pr. Fadwa Ghanimi (ENSCk-UIT)

Année universitaire 2023/2024

End-of-Studies Thesis

For the attainment of the Master's degree in Artificial Intelligence and Virtual Reality

Specialisation: Artificial Intelligence

Host Organization: Open Development & Education

under the theme :

Systematic Reviews and Artificial Intelligence: Mechanisms for Success and Failure

Prepared by:
Mr. TOUIYATE ADNANE

Supervised by:
Mr. BOUJIHA TARIK
(Ibn Tofail University)
Mr. MANSOUR HASSAN
(OpenDevEd)

Defended on September 16, 2024

Academic year 2023/2024

Dedication

This work is dedicated to my beloved family, whose unwavering support, encouragement, and love have been the foundation of all my achievements.

To my parents, for their endless sacrifices and for believing in me even when I doubted myself. Your guidance has shaped who I am today, and your strength has been my greatest inspiration.

To my friends have been my pillars of strength throughout this journey. Your encouragement and understanding have been invaluable.

To my mentors and teachers, Tarik Boujiha, whose wisdom and knowledge have guided me through the complexities of my studies, and Raja Touahni, the coordinator of my master’s program. Your dedication to education has left a lasting impact on my life.

Lastly, to thank all those who have touched my life in ways big and small, helping me reach this milestone.

Thank you.

Acknowledgments

I would like to express my deepest gratitude to all those who have supported me throughout the preparation of this End-of-Studies Thesis for my Master's degree. First and foremost, I would like to extend my sincere thanks to Mr. BOUJIHA TARIK, my academic supervisor, for his invaluable guidance, insightful feedback, and unwavering support throughout this journey. His expertise and encouragement have been instrumental in shaping this thesis. I am also profoundly grateful to Mr. MANSOUR HASSAN, my company supervisor at OpenDevEd, for providing me with the opportunity to undertake my internship at such a reputable organisation. His mentorship, coupled with the practical experience I gained during my time at OpenDevEd, has significantly enriched my understanding of the field. I would like to thank Dr Bethany Huntington and Dr Björn Haßler for providing the opportunity to work alongside them as part of the EEF project, and for their support in working on this thesis. Finally, I would like to thank my family, friends, and colleagues for their constant encouragement and support, which has been a source of strength and motivation. Thank you all for making this journey a fulfilling and memorable one.

Contents

Figures

Abstract

This report is set in the context of a systematic review and meta-analysis that investigates the mechanisms by which educational technology (EdTech) interventions can lead to improved pupil attainment outcomes. The primary objective of the meta-analysis is to identify the key elements that contribute to the success or failure of EdTech interventions. By employing an extensive search methodology, the meta-analysis incorporates advanced techniques such as Latent Dirichlet Allocation (LDA) for topic modelling and AI-assisted screening to enhance the accuracy and efficiency of the review process. The methodology involves systematic searches across multiple databases, including both automated searches using APIs and manual searches in high-relevance databases. A detailed keyword inventory and carefully constructed search queries guide the retrieval of studies, followed by initial deduplication and relevance screening undertaken by human reviewers. The methodology involves systematic searches across multiple databases, including both automated searches using APIs and manual searches in high-relevance databases. A detailed keyword inventory and carefully constructed search queries guide the retrieval of studies, followed by initial deduplication and relevance screening.

WIthin this broader context, this thesis focuses on the use of AI tools. Large Language Models (LLMs) like ChatGPT, Llama, and Mixtral, are used to automate the screening of titles, abstracts, and keywords based on predefined inclusion criteria, reducing manual workload and improving consistency. Manual validation ensures the reliability of AI-assisted screening results.

LDA is employed to classify and validate the literature, providing an additional verification layer to ensure the collected studies' pertinence. The review process concludes with full-text screening, data extraction, and quality appraisal, culminating in a meta-analysis to quantify EdTech interventions' effects. The findings are synthesised into an evidence map, offering a structured overview of existing research and highlighting gaps and overlaps in the literature.

This meticulous approach aims to provide valuable insights into designing and implementing effective EdTech interventions, offering evidence-based recommendations for educators, policymakers, and researchers. The study demonstrates the potential of combining traditional systematic review methods with cutting-edge AI technologies to advance educational research and practice.

Résumé

Ce rapport s’inscrit dans le cadre d’une revue systématique et d’une méta-analyse visant à étudier les mécanismes par lesquels les interventions en technologies éducatives (EdTech) peuvent améliorer les résultats scolaires des élèves. L'objectif principal de la méta-analyse est d'identifier les éléments clés qui contribuent au succès ou à l'échec des interventions EdTech. En employant une méthodologie de recherche approfondie, la méta-analyse intègre des techniques avancées telles que l'Allocation Latente de Dirichlet (LDA) pour la modélisation des sujets et le criblage assisté par l'IA, afin d’améliorer la précision et l'efficacité du processus de revue. La méthodologie implique des recherches systématiques dans plusieurs bases de données, y compris des recherches automatisées via des API et des recherches manuelles dans des bases de données à forte pertinence. Un inventaire détaillé des mots-clés et des requêtes de recherche soigneusement construites guident la récupération des études, suivie d'une déduplication initiale et d'un criblage de pertinence effectué par des réviseurs humains.

Dans ce contexte plus large, cette thèse se concentre sur l'utilisation des outils d'IA. Des modèles de langage à grande échelle (LLM) comme ChatGPT, Llama et Mixtral sont utilisés pour automatiser le criblage des titres, résumés et mots-clés sur la base de critères d'inclusion prédéfinis, réduisant ainsi la charge de travail manuel et améliorant la cohérence. Une validation manuelle garantit la fiabilité des résultats du criblage assisté par l'IA.

La LDA est utilisée pour classifier et valider la littérature, offrant une couche de vérification supplémentaire pour garantir la pertinence des études collectées. Le processus de revue se termine par un criblage des textes intégraux, une extraction de données et une évaluation de la qualité, aboutissant à une méta-analyse pour quantifier les effets des interventions EdTech. Les résultats sont synthétisés dans une cartographie des preuves, offrant une vue d'ensemble structurée des recherches existantes tout en mettant en lumière les lacunes et chevauchements dans la littérature.

Cette approche méticuleuse vise à fournir des informations précieuses sur la conception et la mise en œuvre d’interventions EdTech efficaces, en offrant des recommandations fondées sur des preuves aux éducateurs, décideurs politiques et chercheurs. L'étude démontre le potentiel de combiner des méthodes de revue systématique traditionnelles avec des technologies d'IA de pointe pour faire avancer la recherche et la pratique en éducation.

Chapter 1 : Introduction and General Context

This project report focuses on using AI to support a literature review. To motivate the project work, this chapter presents an introduction to the wider research work and the general context. The wider context is provided by a meta-analysis currently under taken by OpenDevEd, see ↑Haßler et al. (2024). In this chapter, I summarise the key ideas of that meta-analysis.

1.1. Problem statement for the Meta-analysis

The education sector is increasingly using technology to improve student learning. Educational Technology (EdTech) tools have become important in changing how students learn, offering new ways to meet different learning needs. However, it is still crucial to study how well these tools actually help students achieve better results. With so many digital tools and platforms available, educators and policymakers need to figure out which aspects of EdTech are most effective. As education systems around the world aim to keep up with modern learning needs, it’s essential to understand what makes EdTech successful.

Despite the advancements in EdTech, several ongoing problems still affect how well students learn. As schools move towards digital learning, the gap between students with access to technology and those without has widened. Additionally, new technologies are developing so quickly that there isn’t always enough research to show if they really help improve learning outcomes. Often, traditional methods for evaluating these technologies aren’t enough, leading to the use of tools that don’t have strong evidence of their effectiveness. To tackle these issues, it's important to thoroughly understand how EdTech interventions work, including the factors and contexts that lead to their success or failure.

1.2. Challenges and Research Needs in EdTech

In today’s education system, Educational Technology (EdTech) is vital, with key trends including digital tools in classrooms, personalised learning through adaptive technologies, and data analytics to improve learning outcomes. However, implementing these technologies has its challenges.

The digital divide continues to be a formidable obstacle, with inequities in technology access intensifying educational disparities. For instance, an OECD report reveals that students in disadvantaged schools are markedly less likely to have access to high-quality online learning resources compared to their counterparts in more affluent institutions, thereby broadening the achievement gap (↑OECD, 2021). Moreover, the efficacy of educational technology (EdTech) is frequently undermined by insufficient teacher training. A survey conducted by the International Society for Technology in Education (ISTE) highlighted that a mere

16% of educators feel adequately prepared to integrate technology effectively into their teaching practices (↑ISTE, 2020). Additionally, the swift pace of technological advancements often surpasses the available evidence that supports their effectiveness, resulting in the adoption of educational tools that lack robust validation of their impact on learning outcomes (↑Bulman et al., 2016; ↑Vanbecelaere et al., 2023; ↑Escueta et al., 2017).

Given these challenges, it is imperative to conduct thorough research on EdTech interventions. Such research is essential for developing evidence-based practices that can guide decision-making in educational technology, ensuring that investments in EdTech lead to significant and measurable improvements in student learning. Evidence reviews, literature reviews, and meta-analyses play a crucial role in this context, as they enable a systematic evaluation of existing studies and data. This rigorous approach helps in identifying effective strategies, highlighting knowledge gaps, and informing future research directions (↑Spillias et al., 2023; ↑Wang et al., 2022).

Traditionally, these evidence synthesis methods involve manual processes for identifying, selecting, and analysing relevant studies, which can be resource-intensive and time-consuming (↑Nussbaumer-Streit et al., 2021; ↑Wang et al., 2022). However, advancements in artificial intelligence (AI) are beginning to transform research methodologies across various disciplines. AI tools can significantly accelerate tasks such as automated searches and data extraction, enhancing the efficiency and accuracy of evidence synthesis (↑Haßler et al., 2021; ↑Haßler, Hassan & Klune, 2024). Consequently, understanding and leveraging AI capabilities in the research process is crucial for optimising the development and implementation of effective EdTech interventions.

This research is particularly important for ensuring that interventions are effective, scalable, and inclusive, thus addressing the needs of all learners, especially those from disadvantaged backgrounds. Systematic reviews and meta-analyses, such as those conducted by organisations like the Education Endowment Foundation, are pivotal in synthesising existing data to identify effective practices and guide future EdTech implementations, ensuring that educational technologies fulfil their potential to enhance learning outcomes across diverse educational settings.

1.3. Research Question for the Meta-Analysis

Building on the established significance of EdTech, the wider study focuses on a crucial inquiry (↑Haßler et al. 2024): **What mechanisms of EdTech interventions are associated with improved pupil attainment outcomes?**

This research question seeks to uncover the specific processes and elements within EdTech implementations that directly contribute to enhancing student learning and achievement, thereby addressing the core objectives of educational technology.

1.4. Purpose and Significance of the Meta-Analysis

In pursuit of these insights, the purpose of the wider study is to deepen the understanding of effective EdTech practices by identifying and analysing the mechanisms that lead to successful educational outcomes, particularly for disadvantaged communities. Given the rapid adoption of EdTech tools across various educational settings, it is crucial to discern which aspects of these technologies truly benefit student learning and how they can be implemented effectively within the classroom. This research aims to provide educators, policymakers, and EdTech developers with actionable insights into how technology can be most effectively employed in educational contexts to support equitable outcomes. By focusing on the needs of disadvantaged students, the meta-analysis seeks to inform decision-making that prioritises the specific requirements of these communities.

Despite the widespread adoption of EdTech and the considerable body of research affirming its potential to enhance educational outcomes, there remains a significant research gap in understanding the precise mechanisms through which EdTech achieves these outcomes. While it is well-documented that EdTech can work, there is a pressing need to uncover how it works. This includes examining the specific features of EdTech tools, the contexts in which they are most effective, and the ways in which they are integrated into classroom practices. Addressing this gap is essential for ensuring that EdTech implementations are not only effective but also equitable and tailored to meet the needs of all students, especially those from disadvantaged backgrounds.

Ultimately, the wider study contributes to the body of knowledge that supports evidence-based decision-making in education technology, ensuring that investments in EdTech are both impactful and equitable. The findings could inform the design of future technologies and the implementation of educational policies that prioritise student attainment and equity, guiding the effective integration of EdTech within classrooms to enhance learning experiences for all students, especially those in disadvantaged groups.

1.5. Objectives

The wider study conducts a systematic review with meta-analysis. Three primary objectives will be addressed to understand the reasons why EdTech interventions succeed or fail:

- **Identify Core Mechanisms**

Understand the fundamental processes within EdTech interventions that contribute to improved educational outcomes for students.

- **Explore Intermediate Outcomes**

  Investigate the intermediate outcomes associated with increased student attainment and how these outcomes fit within the broader mechanisms of change.

- **Investigate Impact Variability**

  Examine potential differences in the impact of various mechanisms of change within EdTech interventions, with a particular focus on disadvantaged students.

1.6. Overview of the Company

Open Development and Education (OpenDevEd), established in 2014, is a consultancy dedicated to transforming education through a multinational team committed to equity and evidence-based practices. With extensive experience in enhancing teaching and learning across sub-Saharan Africa, the Middle East, Europe, and North America, OpenDevEd is a recognised leader in teacher professional development. The organisation focuses on building local capacity and leveraging technology to achieve impactful and sustainable outcomes. Recently, OpenDevEd has expanded into the UK, supporting marginalised communities and aligning with its global mission to promote equitable education opportunities.

1.6.1. Leveraging Technology for Educational Impact

OpenDevEd utilises advanced technology to enhance educational practices and address key challenges in diverse educational settings. Their approach integrates technology thoughtfully to improve teaching and learning outcomes and support equitable education. Key technological initiatives include:

- **Enhanced Teaching Tools**
  Developing and implementing technology solutions that support personalised learning experiences and adapt to the needs of students across different contexts.

- **Evidence-Based Approaches**
  Utilising data-driven insights to guide the development and implementation

of educational tools and practices, ensuring they are effective and aligned with best practices.

- **Technology for Capacity Building**
  Leveraging technology to build local capacity in educational systems, particularly in low- and middle-income countries, to promote sustainable and impactful educational improvements.

1.6.2. AI-Driven Innovations in Education

As OpenDevEd continues to advance its mission, they are harnessing the power of artificial intelligence (AI) to drive innovation and enhance educational outcomes. Key AI-driven initiatives include:

- **Efficient Evidence Synthesis**
  Employing AI tools to streamline the process of evidence synthesis, such as automated searches and data extraction, to support rigorous research and development of effective EdTech solutions.

- **Adaptive Educational Technologies**
  Developing AI-powered technologies that adapt to real-time feedback and changing educational needs, ensuring that educational tools remain relevant and effective.

1.7. Conclusion

The field of education is at a critical juncture, facing the dual challenge of integrating technology effectively while ensuring equitable access and improving outcomes. AI-driven research methodologies offer transformative potential for understanding and optimising EdTech interventions. The meta-analysis aims to contribute to this transformative wave by providing actionable insights into the mechanisms and impacts of EdTech, ultimately guiding the effective implementation of educational technologies. By addressing the challenges and leveraging innovative solutions, this research seeks to enhance educational practices and outcomes, particularly for disadvantaged students, setting new standards for excellence in education.

Chapter 2: Background

In recent years, the fields of natural language processing (NLP) and machine learning have experienced significant advancements, particularly with the development of Large Language Models (LLMs) and Latent Dirichlet Allocation (LDA). These technologies have revolutionized the processing and analysis of large volumes of text data, offering more efficient and accurate methods for handling information in various domains. This chapter comprehensively reviews existing research on using LLMs and LDA, focusing on their application in systematic reviews, particularly within educational technology (EdTech) interventions. By analyzing these technologies' methodologies, applications, and limitations, this chapter aims to identify gaps in the literature and establish the foundation for the current research's contributions.

3.2 Key Studies and Methodologies

3.2.1 LDA in Systematic Reviews

In their study, ↑Mo et al., 2015 explore the use of Latent Dirichlet Allocation (LDA) to improve the screening phase of systematic reviews. Published in *Systematic Reviews*, the research demonstrates that LDA, by modeling studies as distributions of topics, outperforms the traditional Bag-of-Words (BOW) approach in identifying relevant studies. Using support vector machine (SVM) classifiers, the study found that LDA-based representations achieved higher recall rates, crucial for systematic reviews. The findings suggest that LDA provides a more informative and efficient document representation, helping to reduce the workload of reviewers while enhancing the quality of evidence synthesis.

3.2.2 LLMs in Literature Screening

In their study, ↑Spillias et al., 2023 explore the efficacy of human-AI collaboration in identifying relevant literature for evidence synthesis. Published in *Cell Reports Sustainability*, the research demonstrates how AI, particularly ChatGPT, can be effectively integrated into the systematic review process to enhance accuracy and efficiency. The study involved using ChatGPT to assist in the development of search strings and the screening of articles for a scoping review on community-based fisheries management. It was found that the best outcomes were achieved when the AI was prompted to "reflect" on its responses, and when multiple AI replicates were used. This approach resulted in a high level of agreement with human reviewers. The findings suggest that while AI cannot yet fully replace human input, it significantly improves the speed and reliability of systematic reviews, especially when used in collaboration with human experts.

↑Nakaya et al., 2023 conducted a study to evaluate ChatGPT's ability to classify virtual reality (VR) studies in cardiology, comparing its performance with traditional natural language processing (NLP) methods. The results, published in the *European Heart Journal - Digital Health*, revealed that ChatGPT achieved a classification accuracy of 97%, with high sensitivity and specificity. However, it struggled with studies that did not use head-mounted displays, and its reasoning behind certain classifications was sometimes unclear. The study highlighted the potential of large-scale language models like ChatGPT to facilitate bibliometric analysis, although the authors emphasised the need for further research into the interpretability and reliability of AI-assisted text generation.

3.3 Limitations and Challenges in Existing Research

3.3.1 LDA Limitations

While LDA has been widely adopted for topic modeling in systematic reviews, several limitations persist that impact its effectiveness:

- **Sensitivity to Parameter Settings:** The performance of LDA models can be highly sensitive to the choice of parameters, such as the number of topics and the initial configuration of hyperparameters. This variability can lead to inconsistent results and may require extensive tuning to optimize performance for specific datasets.
- **Interpretability of Topics:** Although LDA provides a compact representation of documents, the interpretability of the resulting topics can still pose a challenge. Topics generated by LDA may be difficult for human reviewers to understand without further enrichment or contextual information, potentially limiting their practical utility in systematic reviews.
- **Domain-Specific Performance:** The study primarily focused on datasets from clinical and social science domains, and the generalizability of LDA-based approaches to other fields remains uncertain. Different domains may require specific adaptations of the model, and the results observed in this study may not directly translate to other systematic review contexts.
- **Computational Complexity:** LDA models, especially when enriched with additional features like multi-word terms, can be computationally intensive, particularly for large-scale systematic reviews. This complexity could limit the scalability and speed of the approach, making it less feasible for real-time or large-scale applications.

3.3.2 LLM Challenges

Despite their advancements, LLMs also face several challenges when applied to the literature screening process:

- **Bias and Accuracy**: LLMs are trained on vast amounts of text data, which can introduce biases in their outputs. These biases can manifest in various ways, such as favoring certain types of studies over others or failing to recognize relevant literature that does not conform to common patterns in the training data. Ensuring the accuracy of LLM outputs requires careful validation and, often, human oversight.
- **Computational Resources**: Training and deploying LLMs is resource-intensive, requiring significant computational power and memory. This can be a barrier for smaller research teams or projects with limited access to high-performance computing resources. Additionally, the ongoing costs of using cloud-based services to run these models can add up, making them less accessible for some researchers.
- **Ethical Considerations**: The use of LLMs raises ethical concerns, particularly related to data privacy and the potential for generating misleading or harmful outputs. When handling sensitive information or making decisions that impact research outcomes, it is crucial to consider the ethical implications of relying on LLMs and to implement safeguards to mitigate potential risks.

3.4. Our Contribution

The current research aims to address several of the gaps identified in the literature:

- **Integrating LDA and LLMs**: This research explores the integration of LDA and LLMs to create a more robust and efficient systematic review process. By combining the strengths of both technologies, the research seeks to improve the accuracy and scalability of topic modeling and literature screening.
- **Enhancing Interpretability**: To tackle the interpretability challenges of LDA, this research investigates the use of N-grams and other preprocessing techniques to produce more coherent and meaningful topics. Additionally, the research aims to develop new methods for validating and refining LDA outputs, making them more actionable for researchers.
- **Optimizing Computational Efficiency**: This research explores strategies for reducing the computational costs associated with LLMs, such as using more efficient model architectures or optimizing the use of cloud-based resources. By making LLMs more accessible, the research seeks to democratize the use of advanced AI tools in systematic reviews.

This chapter has provided a comprehensive review of the state of the art in LDA and LLMs, with a focus on their application in systematic reviews. While both technologies offer significant advantages, they also present challenges that need to be addressed to fully realize their potential. By identifying gaps in the existing literature and proposing targeted contributions, this research aims to advance the field and provide valuable insights for future studies. The next chapter will delve into the methodology used in this research, outlining the specific techniques and tools employed to achieve the research objectives.

Chapter 3 : Conception and Methodology for the meta-analysis

2.1 Conception and Framework of the Meta-analysis

This section introduces the overall conception and the conceptual framework that guides the systematic review and meta-analysis on EdTech interventions (⇡Haßler et al. 2024). The purpose is to explore and identify the mechanisms through which EdTech interventions affect educational outcomes, focusing on rigorous analysis and evidence-based decision-making to optimise future implementations.

2.1.1 Conception of the Meta-analysis

The conception of meta-analysis forth the overarching goals and motivations behind conducting a systematic review and meta-analysis on EdTech interventions. The primary intent is to explore and identify the specific mechanisms through which EdTech interventions influence educational outcomes. The meta-analysis emphasises the necessity for a rigorous examination due to the rapid evolution of EdTech and the increasing reliance on digital solutions in educational contexts. This need became particularly evident during the COVID-19 pandemic, which underscored the critical role of technology in maintaining educational continuity.

The focus is on understanding which components of EdTech are most effective in improving student learning outcomes. By identifying these key mechanisms, the study aims to optimise future EdTech implementations to ensure they deliver equitable benefits for all students, particularly those from disadvantaged backgrounds. This approach is rooted in our commitment to equity, evidence, and accessibility, aiming to bridge the gap between the availability of EdTech tools and their effective use in diverse educational settings.

2.1.2. Conceptual Framework for the Meta-analysis

The study's conceptual framework provides a structured approach to understanding the interplay between EdTech interventions and educational outcomes. It is designed to dissect the complex relationships and processes that underpin successful EdTech implementations. Key elements of the framework include:

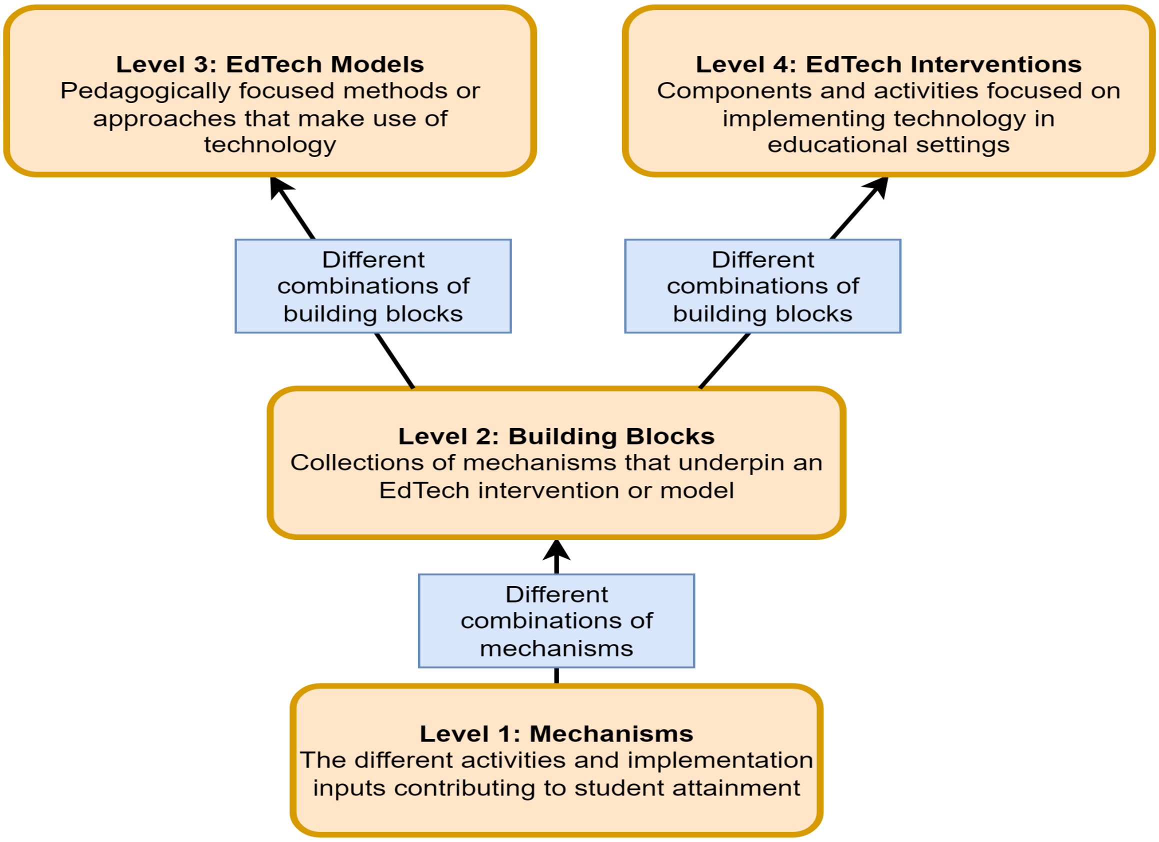
- **Mechanisms of Change:** Identifying the core components and processes within EdTech interventions that drive educational improvements. This includes features such as interactivity, personalization, and data-driven insights.

- **Intermediate Outcomes:** Exploring the intermediary steps and outcomes that link EdTech interventions to improved pupil attainment. These may include enhanced engagement, better retention of information, and increased motivation among students.
- **Contextual Factors:** Examining the environmental and contextual factors that influence the effectiveness of EdTech interventions. This includes socio-economic status, technological infrastructure, and teacher preparedness.
- **Differential Impact**: Investigating how different mechanisms of change impact various student demographics, with a particular focus on disadvantaged pupils. This helps in understanding the equity implications of EdTech implementations.

2.1.3. Conceptual Framework of the Meta-analysis

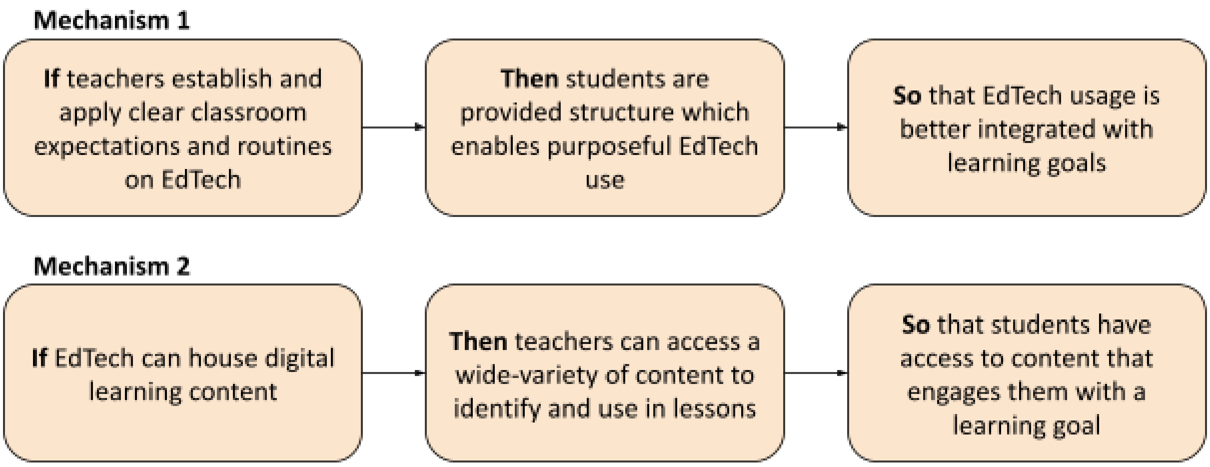
The diagrams serve as visual representations that detail and categorise the theoretical constructs and operational components involved in EdTech interventions. They provide a structured way to visualise the complex relationships and processes hypothesised to influence educational outcomes. Here's the role of each figure:

Figure 2.1. *Conceptualization of EdTech Mechanisms, Building Blocks, Models, and Interventions*



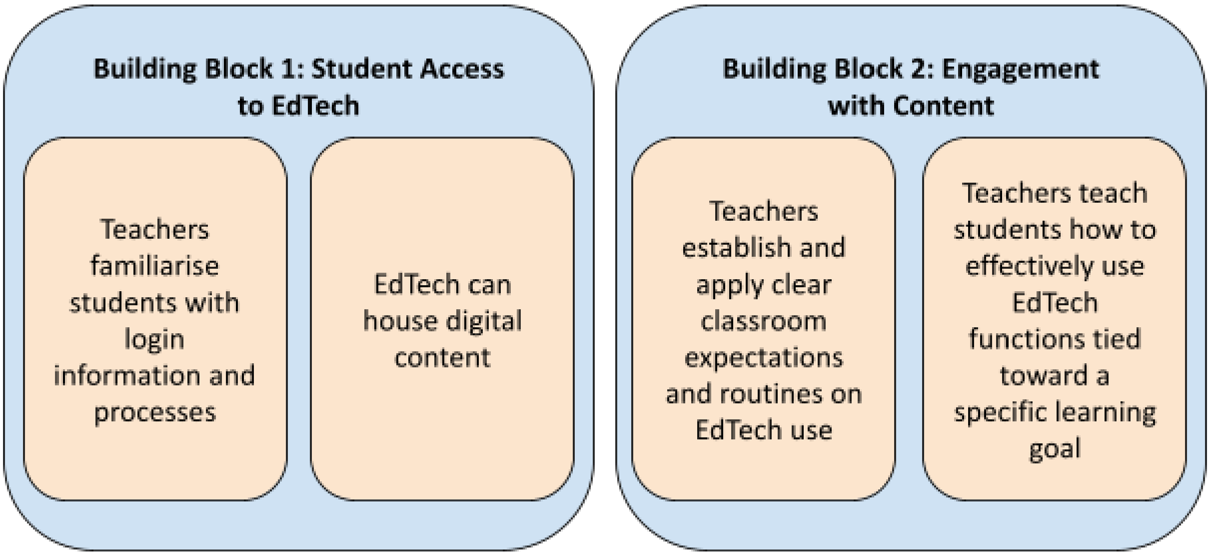
This diagram provides a macro view of how various elements of EdTech interventions are theorised to interact. It visually organises the different layers of intervention — from broad models to specific mechanisms — showing how they collectively contribute to educational outcomes. This aligns with the **Mechanisms of Change** element of the framework by illustrating the overall structure and interaction of key components.

Figure 2.2. *Example of Level 1 Mechanisms*



This diagram drills down into specific mechanisms (Level 1) and illustrates how they function within an "If… then… so" structure. This helps in understanding how individual actions or features of EdTech lead to specific educational outcomes, directly relating to **Intermediate Outcomes** and **Mechanisms of Change**.

Figure 2.3. *Level 2 Building Blocks with Collected Mechanisms*



This diagram maps out how groups of mechanisms (building blocks) work together to support larger EdTech models and interventions. This layering shows the interplay between different mechanisms and how they aggregate to form more complex interventions, highlighting the **Contextual Factors** and **Differential Impact** within the framework.

The diagrams are essential for visually explicating the conceptual underpinnings described in the study's conceptual framework. They not only provide a graphical representation of the theoretical constructs but also aid in systematically breaking down and examining the complex interdependencies within EdTech interventions. This visual breakdown helps in aligning the study's methodologies with its objectives, ensuring a thorough investigation of each component's impact on learning outcomes. By using these diagrams, the study can more effectively communicate the intricate details of how EdTech interventions operate and succeed, thereby supporting evidence-based decision-making and the optimization of EdTech implementations for equitable educational benefits.

For a detailed explanation of the conceptual framework and to view the diagrams, please refer to the protocol for the meta-analysis: ↑Haßler et al. (2024).

2.2. Methodology of the Meta-analysis

This methodology outlines the systematic review process to examine the impact of educational technology (EdTech) interventions on academic attainment among disadvantaged pupils. A review was conducted following best practices for systematic literature reviews in education and EdTech, ensuring a thorough, comprehensive, and unbiased analysis of the existing literature. The methodology presented here follows ↑Haßler et al. (2021) and ↑Haßler et al. (2024).

2.3. Phase 1: Systematic Searches

2.3.1. Automated Cross-Database Search Strategy

To conduct a thorough and comprehensive review, we will employ an automated, cross-database search strategy. This approach is essential because no single database contains the complete set of published materials. Our strategy involves searching multiple databases to access a wider range of sources, increasing the likelihood of identifying relevant studies that may be overlooked by relying on a single database.

2.3.2. Databases, APIs, and CLIs

We will conduct a structured automated search of databases with an Application Programming Interface (API) and a structured manual search of high-relevance databases without API using complex queries. Digital tools, existing software development kits (SDKs), and Command Line Interfaces (CLIs) will be used to perform the structured automated search, interfacing with the database APIs. This approach allows us to repeat searches at regular intervals, supporting the possibility of creating a living review beyond the scope of this project.

2.3.3. Automated Searches

Automated searches are conducted using Application Programming Interfaces (APIs) and Command Line Interfaces (CLIs) provided by major academic databases.

These APIs and CLIs allow for programmatically querying the databases and retrieving relevant articles based on predefined search criteria. The primary tools and databases used in this process include:

- **Google Scholar (scholarly-cli)**: Google Scholar is a widely-used search engine for scholarly literature, encompassing peer-reviewed papers, theses, books, preprints, abstracts, and technical reports across various disciplines. The scholarly-cli tool interfaces with Google Scholar to perform automated searches, collecting comprehensive bibliographic data and metadata. Notably, the scholarly-cli was developed by me, featuring a Python script capable of extracting papers based on specific search parameters. This custom tool enhances our ability to tailor searches precisely to our research needs.
- **OpenAlex (openalex-sdk/cli)**: OpenAlex is an open database that provides extensive coverage of academic publications. The openalex-sdk/cli is used to access and retrieve data from OpenAlex, ensuring that the search captures a wide range of sources, including those not indexed by traditional databases.
- **Scopus (scopus-cli)**: Scopus, provided by Elsevier, is a comprehensive abstract and citation database covering a wide range of academic disciplines. The scopus-cli facilitates structured searches within Scopus, allowing for the retrieval of detailed records and citation data.
- **Web of Science (WoS-cli)**: Web of Science is another major bibliographic database offering comprehensive citation data across multiple disciplines. The WoS-cli is used to perform automated searches, ensuring the inclusion of high-impact articles.

2.3.4. Selected Databases

The selected databases for our review include:

- OpenAlex
- Scopus
- Web of Science
- Scite.ai
- Google Scholar
- British Education Index
- IEEE
- EEF database

- Open Development & Education EdTech evidence library
- J-PAL
- 3ie
- AERDF
- What Works Clearinghouse

2.3.5. Grey Literature

We will include select sources of grey literature, such as the EEF database of education studies and the Open Development & Education EdTech evidence library. These sources are essential as they contain both meta-analyses and literature reviews related to school-based EdTech research. Manual searches will also be conducted on key NGO websites, including J-PAL, 3ie, Advanced Education Research & Development Fund (AERDF), and What Works Clearinghouse.

2.3.6. Keyword Inventory and Search Queries

A comprehensive keyword inventory is developed to guide the search process. This inventory includes terms related to different dimensions of EdTech interventions, categorised into themes such as technology type, educational setting, and target outcomes. Keywords are combined using boolean operators to construct detailed search queries. For example:

```
technology_category AND education_setting_category
```

```
(educational technology OR EdTech OR tablet OR mobile learning) AND (school OR teacher OR classroom)
```

These queries are piloted and refined to ensure they capture relevant studies without generating excessive irrelevant results. The piloting process involves:

- Running initial searches and reviewing the first 100 results for relevance.
- Adjusting search strings based on feedback from manual reviews.
- Documenting all final search strings for transparency and reproducibility.

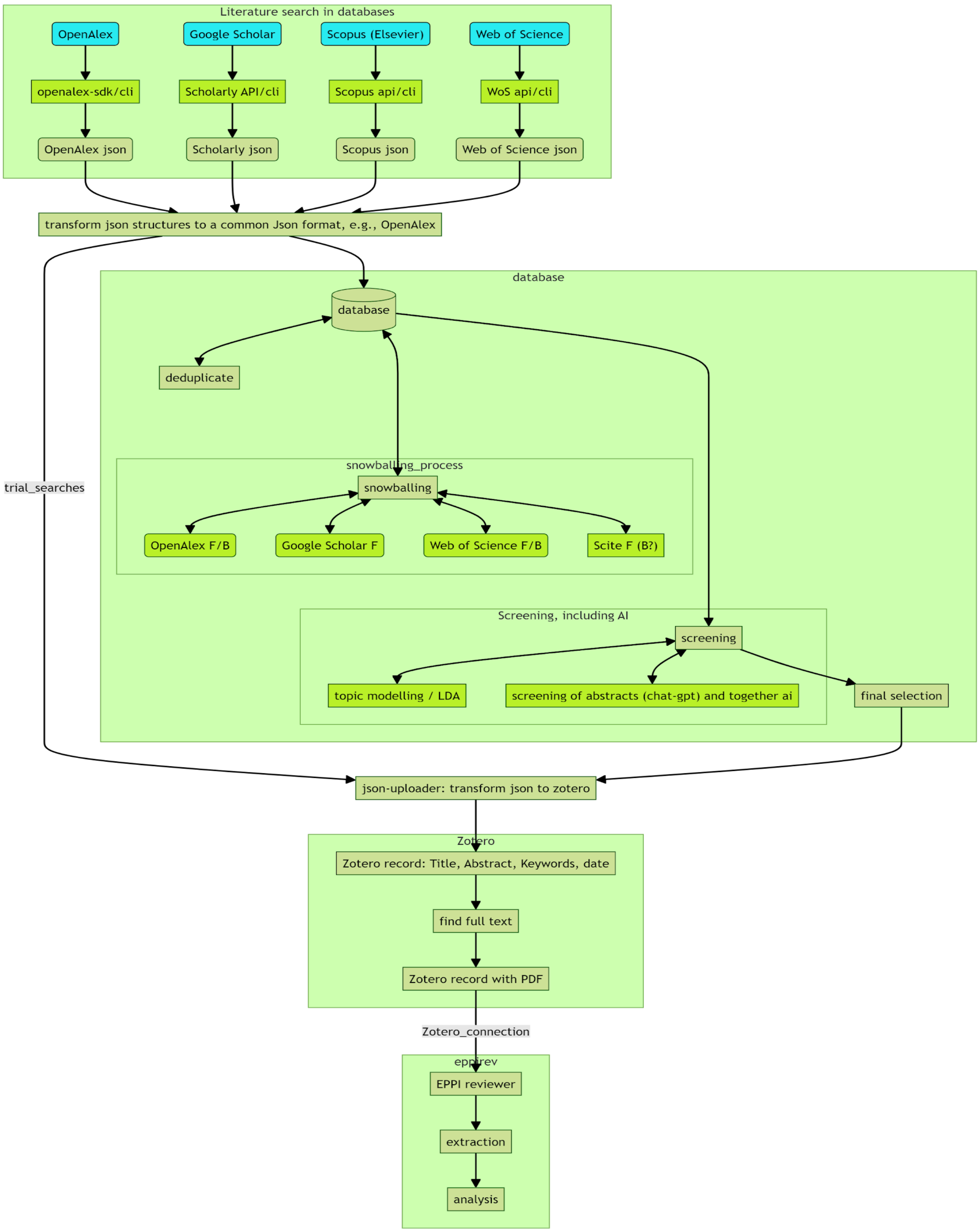
2.3.7. Construction of Search Queries

Search queries will be constructed using a Command Line Interface (CLI) that systematically incorporates the extensive lists of primary keywords within their respective categories. The search queries will use boolean operators to combine relevant keywords within and between categories, ensuring a comprehensive search.

2.3.8. Workflow Diagram for the Meta-analysis

The workflow for conducting the automated searches and subsequent processes is visually represented in the following diagram:

Figure 2.4. *Automated Literature Review Workflow for Systematic Reviews*



The diagram illustrates the workflow from literature search in databases to the final selection and integration with Zotero and EPPI-Reviewer for analysis. It shows the transformation of JSON structures, deduplication, snowballing, and the integration of AI for topic modelling and screening.

2.4. Phase 2 : Validation and classification Using Topic Modelling

2.4.1. Latent Dirichlet Allocation (LDA)

In Phase 2, the systematic review process focuses on rigorously screening and validating the literature identified in Phase 1. This phase is crucial for ensuring the inclusion of only the most relevant and high-quality studies in the review. The screening and validation process involves three primary steps: title, abstract, and keyword screening; validation using Latent Dirichlet Allocation (LDA); and full-text review and quality appraisal.

We will use Latent Dirichlet Allocation (LDA) for topic modelling, a common machine learning approach to identify clusters of similar words within bodies of text. This process will be conducted using readily available libraries in Python, with Zotero used for reference management. Different LDA characteristics will be explored to identify the best-performing characteristics for creating the most relevant topics for our study.

2.4.2. Validation of Articles Using LDA

Having reviewed a large number of publications in the screening, we validated literature against search terms used in our coding tool using Latent Dirichlet Allocation (LDA). This step is not standard in systematic reviews but is an additional step we have added to enhance the validity of our methodology. As we expect a large number of publications to be retrieved, and the data integrity of databases is variable, additional verification of the occurrence of keywords will be undertaken using LDA for topic modelling. This step checks that the papers we have collected align with our intended focus, enhancing the methodological rigour of our review.

The LDA will indicate how publications are characterised, which is helpful for the evidence map. An evidence map provides a structured overview of all the papers that cover the topic, showing overlaps and gaps in the research. LDA can group papers by their identified topics, aiding researchers interested in specific areas. Additionally, LDA may indicate promising areas with sufficient publications to conduct meta-analyses, helping to identify

Topics or areas with a substantial number of papers that might be suitable for subgroup analyses in the meta-analysis stages, thus allowing for a more granular examination of the effects and variations within different subgroups of the research.

2.4.3. Topic List and Refinement

The research team will create a provisional topics list and consult with the EEF to identify any missing topics and ensure the topics modelled are relevant to teaching and learning practices in England. The review team will repeatedly pilot this process until satisfied that a sufficiently broad and relevant set of topics has been captured.

2.5. Phase 3: Eligibility Screening of Titles, Abstracts, Keywords

2.5.1. EPPI-Reviewer for Manual Screening

The long list of studies will be uploaded to EPPI-Reviewer for manual screening, document management, data extraction, and data analysis. Initial deduplication based on DOI will be performed, followed by automated marking of duplicates with a threshold set at 85% similarity. Reviewers will manually check remaining papers with high similarity scores.

2.5.2. Relevance Categories

Two reviewers will independently review the abstracts, coding studies into high, medium, and low relevance categories. High relevance indicates clear alignment with our focus on EdTech and academic attainment, medium relevance is unclear or open to interpretation, and low relevance is clearly unsatisfactory.

2.5.3. Intercoder Agreement

The coding of both reviewers will be compared to explore similarities and conflicts in relevance ratings. Where disagreements occur, a third (senior) reviewer will make the final decision. This process ensures that potentially relevant papers are not excluded prematurely.

2.5.4. Inclusion Criteria

To be included in the review, studies must meet the following criteria:

- Published since 2011
- Peer-reviewed journal articles or select grey literature
- Studies with rigorous causal inference strategies
- Published in English, French, Spanish, or Chinese

- Conducted in countries with high technological readiness
- Focus on students in formal mainstream education in KS1–KS5
- EdTech interventions aimed at improving student academic attainment

2.6. Phase 4 : Additional Automated Screening

2.6.1. Screening with Large Language Models (LLMs)

Our automated approach to screening used several LLMs, including ChatGPT 3.5 from OpenAI, Llama (versions 2 and 3), OpenChat, and Mixtral-7B from Together AI, to screen large numbers of publications (titles, abstracts, and keywords) found in the automated searches. In Phase 1, the automated screening process ran in parallel with (and independently of) screening by human reviewers.

We compiled a list of several databases, including OpenAlex (which also indexes CrossRef), Scopus, and Google Scholar, and selected IGO repositories. These databases are used in a structured automated search with an Application Programming Interface (API) of ChatGPT. We developed software development kits (SDK) to interface with the database APIs. In addition, we created Command Line Interfaces (CLI) to further optimise the interaction with the database APIs and facilitate the execution of our search strategy.

To conduct this search, researchers developed a keyword inventory to search for titles and abstracts of studies. This inventory was used to create a Javascript designed to automate initial searches across the chosen databases. These tools collectively provided an AI-driven review tool that was an effective and user-friendly means of accessing and manipulating data. Furthermore, we thoroughly documented this search strategy and the automated searches, thus enabling repeated searches at regular intervals to provide future updates beyond the life of the project.

In parallel with this, we developed a screening tool consisting of boolean and string questions to feed the AI review tool (↑Haßler et al. 2024). These questions were crafted to ensure the AI could effectively differentiate between relevant and irrelevant publications based on the titles, abstracts, and keywords retrieved in the search.

Boolean Questions: Boolean logic was employed to create precise queries that could narrow down search results by combining or excluding specific keywords. For instance, we used Boolean operators such as AND, OR, and NOT to create questions that required the presence of multiple relevant terms in a study (e.g., "education AND technology AND primary school") or excluded studies that mentioned unrelated or irrelevant topics (e.g., "NOT medical"). The aim was to ensure that the AI focused on publications that were highly likely to align with our research focus, minimising the retrieval of unrelated papers.

String Questions: String-based questions were used to identify specific phrases or combinations of words within the titles and abstracts of publications. These questions were designed to capture nuances in language that might indicate the relevance or irrelevance of a paper. For example, questions could target specific educational interventions, methodologies, or outcomes by searching for exact phrases like "digital learning tools" or "student achievement." This allowed the AI to prioritise papers that discussed key concepts central to our review.

Rationale: The development of these questions was guided by the need to achieve high specificity and sensitivity in the automated screening process. By using Boolean logic, we could craft complex queries that combined multiple conditions, ensuring that only the most pertinent studies were flagged for further review. String questions, on the other hand, allowed us to capture the exact terms and phrases that reflected the specific focus of our systematic review.

These questions underwent multiple iterations, with researchers manually assessing the AI's responses and refining the questions to improve accuracy. The process of trial and error was crucial in fine-tuning the AI's ability to identify relevant studies while reducing false positives and negatives. This iterative approach ensured that the AI screening tool became increasingly reliable over time, effectively complementing the human review process and enhancing the efficiency and accuracy of our systematic review.

A vital part of this process was the trial-and-error feedback loops between human researchers and AI tools in formulating valid and reliable screening questions for API-driven automated searches. The importance of manually developing rigorous screening criteria and questions to be used with APIs cannot be understated.

2.6.2. Title and Abstract Screening

The purpose of the title and abstract screening is to systematically review the papers identified in the automated searches to determine their relevance to the meta-analysis study (↑Haßler et al. 2024). This involves developing specific screening questions based on our inclusion criteria to guide both human reviewers and LLMs in the screening process.

For example, one inclusion criterion for ‘research design’ is:

Studies with rigorous causal inference strategies, including randomised controlled trials and quasi-experimental methods (e.g., instrumental variables, differences-in-differences, fixed effects, regression discontinuity, propensity score matching). Mixed-methods studies that include a quantitative element with a rigorous causal inference strategy. Screening questions developed for this criterion include:

- **research_approach_rct**: Is the study design a randomised control trial (RCT)?
- **research_approach_qed**: Does the study utilise a quasi-experimental design (QED), such as difference-in-difference or other types of quasi-experimental designs?
- **research_approach_impact_evaluation**: Is the study classified as an impact evaluation (using methods to rigorously assess the causal impact of an intervention, policy, or programme)?
- **research_approach_process_evaluation**: Is the study classified as a process evaluation (systematically assesses the implementation processes, mechanisms, and contextual factors involved in delivering an intervention, policy, or programme)?
- **research_approach_qualitative**: Does the study predominantly employ qualitative methodologies?
- **research_approach_mixed_methods**: Does the study employ a mixed-methods approach?
- **research_approach_literature_review**: Is the study classified as a literature review?
- **research_approach_systematic_review**: Is the study classified as a systematic literature review?
- **research_approach_meta_analysis**: Is the study classified as a meta-analysis?

These questions help determine if the study meets our inclusion criteria. For example, a study marked "yes" for RCT would be relevant, while a study marked "yes" for a literature review would not and could be excluded. This process is applied to all points of the inclusion criteria, except for those where information can be obtained from basic Zotero data, such as the year of publication.

Additional string questions were included for the AI screening to capture any incorrect boolean answers and collect useful information for evaluating the AI screening.

2.6.3. Evaluation and Iteration

The evaluation and iteration phase involves a systematic approach to refining the AI-assisted screening process. Initially, the screening tool, designed for manual review, is fed into the AI review tool. This screening tool is tested on a small subset of publications (up to 50 per iteration) to assess the preliminary performance of the AI. Researchers manually evaluate the AI's responses, ensuring accuracy and reliability. Discrepancies between the AI outputs and human evaluations are closely analysed, and adjustments are made to optimise the AI's performance.

2.6.4. Identification of Highest-Performing LLMs

A critical part of this phase is the identification of the highest-performing Large Language Models (LLMs) from the set initially tested. This process begins with an in-depth performance evaluation of the LLMs, including ChatGPT 3.5, Llama (versions 2 and 3), OpenChat, and Mixtral-7B from Together AI. The performance metrics include:

- **Accuracy:** The ability of the model to correctly classify studies as relevant or irrelevant based on the screening criteria.
- **Consistency:** The model's ability to provide stable and repeatable results across different datasets.
- **Precision in Inclusion/Exclusion:** The ability to minimise false positives (incorrect inclusion) and false negatives (incorrect exclusion).
- **Processing Time**: The efficiency of the model in terms of the time taken to screen and classify studies, which is crucial for large datasets.

After evaluating the performance of various models, we identified the two highest-performing models: **Mixtral-8x22B Instruct v0.1** and **Mixtral-8x7B Instruct v0.1**. These models were selected based on their superior accuracy, consistency, and precision in classifying relevant and irrelevant studies. The Mixtral models outperformed others in handling complex queries, demonstrating a high capability in reducing false positives and false negatives. By employing these two models, we were able to enhance the reliability of the automated screening process, ensuring that the included studies were highly relevant to our research objectives. The rigorous selection process involved extensive testing and validation, confirming that these models were the most effective for our systematic review.

The two highest-performing models are selected based on these metrics. These models will then undergo further testing with a more extensive dataset.

2.6.5. Full-Scale Screening and Question Refinement

After identifying the top two models, these LLMs are used to rerun the screening questions on the full set of papers that have been manually screened. This step serves two purposes:

- **Validation:** To ensure that the AI models can replicate the decisions made by human reviewers when applied to a complete dataset.
- **Optimization:** To identify any further adjustments needed in the screening questions or model parameters to improve accuracy.

During this phase, researchers compare the AI's results with the manual screening decisions across the entire dataset. When discrepancies arise between the AI's classification and the manual review, an informal analysis is conducted. The focus is on understanding why these differences occurred and making the necessary adjustments to the screening questions or model parameters to better align the AI's decisions with human judgement.

2.6.6. Choosing Questions for Final Screening

Once the full-scale testing is complete, the screening questions are refined to maximise their effectiveness. This involves:

➢ **Selecting Initial Questions:** The first step is to choose an initial set of screening questions that will guide the final decisions for including or excluding studies. These questions are selected based on their potential to accurately determine relevance. For example, given our interest in Randomised Controlled Trials (RCTs) and quasi-experimental designs (QEDs) that focus on EdTech interventions involving school students, an example set of questions might include:

  - **Is the study a Randomised Controlled Trial (RCT)?**
  - **Is the study a Quasi-Experimental Design (QED)?**
  - **Does the study involve the use of educational technology (EdTech)?**
  - **Does the study focus on school students?**

  If the answers to these questions are "YES," the study is included. If the answers are "NO," the study is excluded. These questions help ensure that the selected studies align with the specific criteria of our systematic review.

➢ **Iterative Refinement:** The selected questions are then subjected to iterative testing with the LLMs. During this phase, the primary concern is assessing the accuracy of the chosen questions in determining whether studies should be included or excluded. Instead of refining the wording of the questions (which was addressed during the pilot stage), this step focuses on testing different combinations of the 3-5 key questions to evaluate their collective accuracy in screening. The goal is to determine whether the chosen combination of questions consistently yields correct inclusion or exclusion decisions. Through this process, we can identify the most effective combination of questions and refine our approach to maximise accuracy.

- ➢ **Identifying Key Questions:** As the questions are tested and refined, those that consistently lead to accurate include or exclude decisions are prioritised. These key questions become the foundation of the final screening process. The focus here is on identifying the optimal combination of questions that most reliably differentiates between relevant and irrelevant studies.
- ➢ **Final Selection and Validation:** The refined and prioritised questions undergo further testing to confirm their reliability and validity. Only those questions that consistently perform well in terms of accuracy are chosen for the final automated screening process. This ensures that the final set of questions is robust and effective in screening large volumes of literature.

These final questions are then used in the automated screening process, ensuring that only the most relevant studies are included while irrelevant ones are effectively excluded.

2.6.7. Integration and Final Moderation

The outcomes of the automated and manual screenings are compared and moderated to ensure consistency. In this final moderation phase, human researchers, who make the ultimate inclusion or exclusion decisions, review any discrepancies identified between the AI-assisted screening and the manual process. This comparison is not just about resolving differences but also involves assessing the performance and accuracy of the AI screening relative to human judgement. By analysing how well the AI aligns with human decisions, we can gain insights into the effectiveness of the AI tools in replicating human decision-making in systematic reviews. This dual approach, combining manual and AI screening, aims to streamline the screening process, enhance efficiency, and reduce potential bias, while maintaining rigorous standards. The iterative feedback loops between human and AI evaluations are key to refining the tools and ensuring they align with the study's rigorous methodological standards. The comparison between AI and human screening accuracy also offers valuable lessons for the future of systematic reviews in research, particularly in understanding the strengths and limitations of AI in this context.

2.7. Phase 5: Creating the Evidence Map

2.7.1. Role of Latent Dirichlet Allocation (LDA) in Evidence Mapping

In Phase 5, we use Latent Dirichlet Allocation (LDA) to create an evidence map that visually represents and organises the evidence from the included studies. This phase aims to identify patterns, gaps, and clusters in the data, providing a clear overview of the landscape of EdTech interventions and their mechanisms.

2.7.2. Evidence Mapping Process

Data Collection and Preparation:

- **Collect Data**: Gather the full text of all included studies after the screening process.
- **Preprocess Data**: Clean and preprocess the text data by removing stop words, stemming, and tokenizing the text.

Apply LDA:

- Run LDA Analysis: Apply LDA to identify underlying topics within the collected studies. For example, we might choose to identify 10 topics.
- Topic Examination: Each topic is represented by a set of words that co-occur frequently in the studies. For example:
    - Topic 1: {online, platform, learning, student, access, resource}
    - Topic 2: {interactive, tool, engagement, activity, classroom, participation}
    - Topic 3: {assessment, test, score, performance, evaluation, improvement}
    - Topic 4: {video, tutorial, content, student, learning, multimedia}
    - Topic 5: {gamification, game, motivation, student, engagement, fun}

Verification and Characterization:

- Keyword Validation: Verify that the key terms you expected are adequately represented within the identified topics. This step ensures that the topics are relevant to the research question.
- Characterization of Topics: Analyse the distribution of topics across the studies to understand how different EdTech interventions and outcomes are represented.

Visualisation:

- Create Evidence Map: Develop visual representations such as heat maps, bubble charts, or network diagrams to show the distribution and relationships between topics. For instance, a heat map might show that interactive tools are frequently associated with student engagement and improved test scores, while online platforms have mixed results.

Identify Patterns and Gaps:

- Cluster Identification: Identify clusters of studies within the same topic. For example, you might find a significant cluster of studies focused on interactive tools and their impact on student engagement.
- Gap Analysis: Identify areas where there is a lack of research, highlighting potential areas for future studies.

Integration with Meta-Analysis:

- Promising Areas for Meta-Analysis: Use the evidence map to identify topics with sufficient studies for meta-analysis. For instance, if there is a substantial number of studies under the topic "interactive tools and student engagement," this could be a promising area for further quantitative analysis.

2.7.3. Using the Evidence Map to Identify Successful EdTech Interventions

The evidence map serves as a crucial tool to understand and identify which specific EdTech interventions are responsible for student success by:

- Highlighting Effective Interventions:
  - Visualisation of Success Patterns: The evidence map visually highlights which EdTech interventions (e.g., interactive tools, online platforms) are consistently associated with positive student outcomes such as improved engagement and higher test scores.
  - Cluster Analysis: By identifying clusters of studies that report similar outcomes, we can see which interventions show consistent success across different contexts and studies.
- Understanding Contextual Factors:
  - Contextual Insights: The evidence map can reveal how the effectiveness of different EdTech interventions varies based on contextual factors such as the type of educational setting (e.g., urban vs. rural schools), age group of students, and specific implementation strategies.

    - Tailored Recommendations: This allows for more tailored recommendations on which EdTech interventions are likely to be successful in specific contexts.
- Identifying Research Gaps:
    - Highlighting Under-Researched Areas: The evidence map can identify gaps in the research, indicating areas where more studies are needed.
    - Guiding Future Research: Researchers can focus on these gaps to build a more comprehensive understanding of EdTech interventions across different contexts.
- Supporting Meta-Analysis:
    - Quantitative Synthesis: The evidence map helps in identifying areas with sufficient studies for meta-analysis, enabling a quantitative synthesis of the data.
    - Effect Size Calculation: By calculating the effect sizes for various interventions within the evidence map, we can quantitatively assess which interventions have the most significant impact on student attainment.

2.8. Conclusion

The systematic review methodology integrates automated and manual processes to ensure a comprehensive and unbiased analysis of EdTech interventions (↑Haßler et al. 2024). By combining structured searches, topic modelling, manual screening, and AI-based tools, we aim to provide a robust and thorough review of the existing literature, identifying the mechanisms and building blocks that underpin successful EdTech interventions for disadvantaged pupils. The evidence map created using LDA plays a pivotal role in visually representing the data, identifying effective interventions, understanding contextual factors, highlighting research gaps, and supporting meta-analyses to draw meaningful conclusions about what works in EdTech and why.

This thesis will focus specifically on the latter aspects: The use of LDA and LLM for systematic reviews and meta-analysis, responding to challenges set out in ↑Haßler, Hassan & Klune (2024).

Chapter 4 : Using LDA for meta-analysis

As we have seen in the previous chapters there are significant opportunities for the use of artificial intelligence and other computational methods in the field of meta-analysis. This chapter presents the focus on the current project. The research question for the Masters thesis is:

- How can LDA and LLM contribute to a meta-analysis on EdTech for disadvantaged children?

4.1. Introduction

This chapter details the practical implementation of the methodologies and tools used in this research project. We begin by exploring the **Scholarly-cli** tool, which streamlines data collection from Google Scholar, facilitating systematic reviews and meta-analyses. This is followed by an in-depth look at Latent Dirichlet Allocation **(LDA)** for topic modelling, where we discuss the implementation of N-grams to improve topic coherence and relevance. Finally, we delve into the use of Large Language Models **(LLMs)** from Together AI, focusing on how these advanced AI models were employed to extract structured information from academic papers. Throughout this chapter, the emphasis is placed on the technical processes, tools, and evaluation metrics that ensure the reliability and accuracy of our research outcomes.

4.2. Scholarly-cli: Data Collection

In this section, we explore the process of data collection using the `scholarly-cli` tool, a command-line interface designed to interact with Google Scholar. This tool allows researchers to efficiently search for scholarly articles and retrieve them in various formats such as JSON, BibTeX, or TSV.

4.2.1. Overview of Scholarly-cli

`Scholarly-cli` streamlines the process of conducting systematic literature reviews or meta-analyses by enabling advanced searches on Google Scholar. The tool supports various search parameters, including limiting the number of results, sorting by relevance or date, and retrieving specific metadata like patents and citations.

4.2.2. Installation and Setup

The `scholarly-cli` tool is built using Python and requires the installation of several packages, including `scholarly` and `ProxyGenerator`. For enhanced functionality, such as bypassing rate limits, it can be configured to use a proxy service like Scraper API.

After installing the necessary packages and setting up the tool, users can configure it with an API key to access additional features provided by the proxy service.

4.2.3. Running the Script

Once the tool is installed and configured, users can run the `scholarly-cli` script with specific commands to collect data. Below are examples of how to run the script:

Basic Search: To search for publications related to "climate change" and "Africa" and retrieve the first 20 results:

```
scholarly-cli search "climate change" AND Africa --limit 20 --json
--save scholarly_data
```

This command will save the results in a JSON file named scholarly_data.json.

Using Date Ranges: To search within a specific date range, for example, between 2010 and 2020:

```
scholarly-cli search "climate change" AND Africa --date 2010-2020
--limit 20 --json --save scholarly_data
```

Output in Different Formats: To save the results in BibTeX format:

```
scholarly-cli search "climate change" AND Africa --limit 20
--bibtex --save scholarly_data
```

Using API Key with Proxies: If you have set up a Scraper API key, you can configure the script to use it:

```
scholarly-cli config
```

- After configuring, the script will use the proxy service to enhance data retrieval.

4.2.4. Handling Proxies and API Keys

To avoid hitting Google Scholar's rate limits, `scholarly-cli` can be configured to use proxies. Users can input their API key through a simple configuration command, enabling seamless integration with proxy services like Scraper API. This setup enhances the tool's ability to retrieve data without interruptions.

4.2.5. Testing and Validation

The `scholarly-cli` tool was rigorously tested using various search queries and configurations. The results demonstrated the tool's reliability and effectiveness in collecting scholarly data, making it a valuable asset for researchers.

4.2.6. Source code

For detailed information and access to the source code, you can visit the public GitHub repository: scholarly-cli.

4.3. Overview of LDA

Latent Dirichlet Allocation (LDA) is a fundamental tool in topic modelling, which allows us to uncover hidden thematic structures within a large corpus of text. In this section, we detail the implementation of LDA in our project, the validation of the topics generated, the creation of an Evidence Map, and the use of N-grams to enhance the quality of the topics.

4.3.1. Implementation of LDA

In our implementation, the LDA model was applied to a collection of academic papers, with the goal of identifying underlying topics that span across different documents. The process involved several key steps:

- **Data Preprocessing**: Initially, the text data was preprocessed to remove noise and irrelevant information. This included tokenization, removal of stopwords, and lemmatization. This step ensured that the input to the LDA model was clean and meaningful.

- **Incorporating N-grams**: To capture more contextually meaningful phrases, we extended our preprocessing pipeline to include N-grams (specifically 2-grams and 3-grams). N-grams are contiguous sequences of words that often carry more semantic weight than individual words. For example, "machine learning" as a 2-gram conveys more specific information than treating "machine" and "learning" separately.

    - **Without N-grams**: The LDA model would treat each word independently, potentially missing out on important context provided by word combinations.

    - **With N-grams**: By incorporating 2-grams and 3-grams, the model could better capture context, leading to more coherent and specific topics.

Here's how the preprocessing was implemented:

```
def preprocess(text):
    tokens = word_tokenize(text.lower())
    tokens = [word for word in tokens if word.isalpha()]
    tokens = [word for word in tokens if word not in stop_words]
    tokens = [lemmatizer.lemmatize(word) for word in tokens]

    # Generate 2-grams and 3-grams
    bigrams = generate_ngrams(tokens, 2)
    trigrams = generate_ngrams(tokens, 3)

    # Return bigrams first, then trigrams
    return bigrams + trigrams
```

- **Model Training**: The training of the Latent Dirichlet Allocation (LDA) model is a crucial step in the topic modelling process, as it determines the quality and coherence of the topics generated. This process involves several key parameters and iterative procedures that ensure the model accurately captures the underlying structure of the text data.
    - **Data Preparation for Training :** Before training the LDA model, the text data undergoes preprocessing to remove noise and ensure consistency. After preprocessing, each document in the corpus is converted into a bag-of-words (BoW) representation. This representation is a list of tuples where each tuple consists of a word ID and the word's frequency in the document. The BoW representation is essential for LDA, as it forms the input on which the model operates.

    ```
    corpus = [dictionary.doc2bow(text) for text in
    data['tokens']]
    ```

Key Parameters in LDA Model Training

- **Number of Topics (num_topics)**: The num_topics parameter specifies the number of topics the model should generate. This is a critical decision, as too few topics might oversimplify the data, while too many topics might lead to redundancy and overfitting. For our analysis, we set num_topics to 15, based on prior knowledge of the domain and exploratory analysis.

```
num_topics = 15
```

- **Alpha (alpha)**: The alpha parameter controls the distribution of topics in a document. A higher alpha value results in documents that are mixtures of many topics, while a lower alpha leads to documents being dominated by a few topics. In our implementation, we set alpha to 1, aiming for a balanced distribution where each document can be associated with multiple topics.

```
alpha = 1
```

- **Eta (eta)**: The eta parameter influences the distribution of words in topics. A higher eta value means topics are likely to be composed of many words, while a lower eta value leads to more specific topics with fewer, more concentrated words. Like alpha, eta was set to 1 to allow a balanced distribution of words across topics.

```
eta = 1
```

- **Number of Passes (passes)**: The passes parameter defines how many times the model goes through the entire corpus during training. More passes allow the model to better adjust its internal parameters, leading to a more refined topic distribution. In our case, the model was trained with 17 passes, which provided a good balance between computational efficiency and model accuracy.

```
passes = 17
```

The Training Process

Once the parameters were set, the LDA model was trained using the preprocessed corpus. The training process involved the following steps:

- **Initialization**: The model starts with a random assignment of topics to words. Each document is assumed to be a mixture of topics, and each topic is a mixture of words.

- **Gibbs Sampling**: LDA uses a technique called Gibbs Sampling to iteratively refine the assignment of topics to words. During each iteration, the model updates the topic assignment for each word in each document based on the current state of the model. This process gradually improves the model's understanding of the underlying topic structure.

- **Convergence**: Over multiple passes through the corpus, the model's parameters (the topic-word and document-topic distributions) converge to a stable state. This convergence indicates that the model has found a good approximation of the latent topics within the data.

```
lda_model = LdaModel(corpus, num_topics=num_topics,
id2word=dictionary, passes=passes, alpha=alpha, eta=eta)
```

- **Evaluation**: After training, the model's performance was evaluated using metrics like Perplexity and Coherence Score. Perplexity measures how well the model predicts a sample of unseen data, with lower values indicating a better model. Coherence Score evaluates the interpretability of the topics by measuring the semantic similarity of words within each topic.

```
perplexity = lda_model.log_perplexity(corpus)
coherence_model_lda = CoherenceModel(model=lda_model,
texts=data['tokens'], dictionary=dictionary, coherence='c_v')
coherence_lda = coherence_model_lda.get_coherence()
print(f'Perplexity: {perplexity}')
print(f'Coherence Score: {coherence_lda}')
```

- **Model Validation**: After training, the LDA model's output was validated by mapping the generated topics to predefined thematic areas relevant to our research. This step ensured that the topics were not only statistically significant but also aligned with the research objectives. Validation was achieved by comparing the LDA output with a set of predefined topics and filtering out those that did not meet a certain probability threshold.

```
def validate_topics(doc_topics, predefined_topics):
    validation_results = []
    for topics in doc_topics:
        valid_topics = {topic: prob for topic, prob in
topics.items() if topic in predefined_topics}
        validation_results.append(valid_topics)
    return validation_results
```

4.3.2. Differences with and without N-grams

The inclusion of N-grams significantly impacted the quality and coherence of the topics generated by the LDA model. Without N-grams, the model might produce topics that are less specific, as it relies on individual words that may not convey enough context. By incorporating 2-grams and 3-grams, the model could identify more nuanced and contextually rich topics. For example, a topic about "digital literacy" is more informative as a 2-gram than as two separate words "digital" and "literacy."

In practice, the use of N-grams led to more focused and meaningful topics, particularly in domains where specific phrases and terminologies are common. This approach also helped in reducing ambiguity by combining words that are frequently used together in the literature.

4.3.3. Creating the Evidence Map

Following the validation of topics, we visualised the distribution of these topics across the corpus through an Evidence Map. The Evidence Map serves as a strategic tool to highlight the presence, prominence, and gaps in research topics across different documents.

- **Bubble Chart Visualisation**: A bubble chart was used to represent the distribution of topics across documents. The size of each bubble reflects the prominence of the topic within a document, providing a quick visual cue about the importance of each topic.

```
bubble_plot = sns.scatterplot(data=bubble_data_long, x='Topic', y='Paper', size='Size', hue='Topic', sizes=(20, 2000), legend=None, alpha=0.6)
```

- **Heatmap Visualisation**: A heatmap provides a comprehensive overview of topic distribution across the entire corpus. The intensity of the colour in the heatmap indicates the strength of a topic within a document, which is particularly useful for identifying areas with concentrated research or gaps.

```
sns.heatmap(heatmap_data, cmap="YlGnBu", cbar=True)
```

The evidence maps, generated from the LDA model with N-grams, offered a clear visual summary of the research landscape, enabling us to identify key areas for further analysis and potential meta-analyses. This approach ensured that our topic modelling was not only statistically rigorous but also practically relevant, providing actionable insights into the corpus of academic papers.

4.4. Conclusion

The implementation strategies outlined in this chapter underscore the importance of using advanced tools and methodologies to achieve accurate and reliable research outcomes. By leveraging the **Scholarly-cli** tool, we efficiently collected and managed scholarly data, setting a strong foundation for the subsequent analyses. The use of **Latent Dirichlet Allocation (LDA)**, enhanced by N-grams, provided deep insights into the thematic structures of the research corpus.

Chapter 5 : Using LLM for meta-analysis

This chapter presents an overview of LLMs, and how to use the models from Together AI.

5.1. Introduction to LLM and Together AI

Large Language Models (LLMs) are advanced AI models that have been trained on vast amounts of text data to understand and generate human-like text. These models are capable of performing a wide range of natural language processing tasks, such as text summarization, translation, question-answering, and more. LLMs have gained prominence due to their ability to understand context and generate coherent, contextually relevant text, making them invaluable tools for a variety of applications, including research, customer service, and content creation.

Together AI is a platform that provides access to various LLMs, including models from well-known AI developers such as Meta, OpenAI, and others. Together AI offers a range of models, categorised by their capabilities and intended use cases. One of the key categories of models provided by Together AI is **instruct models;** these models are specifically fine-tuned to follow human instructions more effectively, making them ideal for tasks that require precise and structured outputs.

In this implementation, we focused on using instruct models provided by Together AI to perform a specific task: extracting structured information from academic papers. The instruct models are particularly well-suited for this task because they are designed to follow detailed prompts and produce outputs that adhere to specific formats, such as JSON.

5.2. Model Selection

The following instruct models were selected for evaluation:

- **Mixtral, Version 1 (Low Parameters)**:
    - **Alias**: m1
    - **Model**: MISTRALAI/MIXTRAL-8X7B-INSTRUCT-V0.1
    - **Temperature**: 0.0
    - **Max Tokens**: 10,000
    - **Cost per Million Tokens**: $0.60
- **Mixtral, Version 1 (High Parameters)**:

- **Alias**: m2
- **Model**: mistralai/Mixtral-8x22B-Instruct-v0.1
- **Temperature**: 0.0
- **Max Tokens**: 10,000
- **Cost per Million Tokens**: $1.20

- **LLAMA3**:
  - **Alias**: L2
  - **Model**: meta-llama/Llama-3-70b-chat-hf
  - **Temperature**: 0.0
  - **Max Tokens**: 5,000
  - **Cost per Million Tokens**: $0.90
- **OpenChat 3.5**:
  - **Alias**: c1
  - **Model**: openchat/openchat-3.5-1210
  - **Temperature**: 0.5
  - **Max Tokens**: 2,600
  - **Cost per Million Tokens**: $0.20

These models were chosen because they are well-suited for the task of extracting information from text and formatting it into a structured format like JSON. The instruct models, in particular, are designed to better understand and follow complex instructions, making them ideal for tasks that require high accuracy and precision.

5.3. The Prompt

The models were provided with a specific prompt designed to extract detailed information from the titles and abstracts of academic papers. The prompt was structured as follows:

```
Task: Extract specific information from the provided paper's title and
abstract and format the results as a JSON object.
```

```
Required JSON Attributes:

- topic_language (boolean): Is the abstract written in English?
- research_approach_quantitative (boolean): Does the study predominantly
employ quantitative methodologies?
- research_approach_rct (boolean): Is the study design a randomised control
trial (RCT)?
- research_approach_qed (boolean): Does the study utilise a
quasi-experimental design (QED)?
- research_approach_impact_evaluation (boolean): Is the study classified as
an impact evaluation?
- research_approach_process_evaluation (boolean): Is the study classified as
a process evaluation?
- research_approach_qualitative (boolean): Does the study predominantly
employ qualitative methodologies?
- research_approach_mixed_methods (boolean): Does the study employ a
mixed-methods approach?
- research_approach_literature_review (boolean): Is the study classified as
a literature review?
- research_approach_systematic_review (boolean): Is the study classified as
a systematic literature review?
- research_approach_meta_analysis (boolean): Is the study classified as a
meta-analysis?
- research_design_primary_focus_type (boolean): Does the study employ
rigorous causal inference strategies?
- topic_health (boolean): Is the study focused on health or health
education?
- topic_climate_change (boolean): Is the study focused on education related
to climate change?
- topic_ed_sust_dev (boolean): Is the study focused on education for
sustainable development?
- topic_education (boolean): Is the study focused on education for school
students?
- topic_edtech (boolean): Is the study focused on EdTech or educational
technology?
- topic_disadvantaged (boolean): Does the study specifically address
disadvantaged students?
- topic_ses (boolean): Is the study specifically focused on students with
low socioeconomic status (SES)?
- topic_send (boolean): Is the study specifically focused on students with
disabilities or special educational needs (SEND)?
- topic_minorities (boolean): Is the study specifically focused on students
```

```
who belong to an ethnic minority?
- topic_minorities_black (boolean): Is the study specifically focused on
students of Black ethnicity?
- topic_minorities_caribbean (boolean): Is the study specifically focused on
students of Caribbean ethnicity?
- topic_free_school_meals (boolean): Is the study specifically focused on
students eligible for free school meals (FSM)?
- topic_looked_after (boolean): Is the study specifically focused on "looked
after" students?
- topic_low_income (boolean): Is the study specifically focused on students
from low-income families?
- demographic_preschool_students (boolean): Is the study focused on
preschool students?
- demographic_school_students (boolean): Does the study primarily examine
the experiences of school students?
- demographic_students_tertiary (boolean): Is the study focused on students
in tertiary education?
- population_type_disadvantage (boolean): Does the study focus on
disadvantaged students in formal education?
- geographic_focus_techready (boolean): Does the study focus on countries
with high technological readiness?
- intervention_type (boolean): Does the study examine interventions using
digital approaches?
- intervention_barriers (boolean): Does the abstract mention barriers to the
use of educational technology?
- parents (boolean): Is the study specifically focused on parents?
- teachers (boolean): Is the study specifically focused on teachers?
- subject_math (boolean): Does the study focus on mathematics?
- subject_lit (boolean): Does the study focus on literacy?
- learning_outcomes (boolean): Does the study investigate learning outcomes
or academic achievement?
- outcomes_types (boolean): Does the study use standardised tests to measure
student attainment?
- topic_subject (string): List the specific school subjects focused on in
the study.
- education_level (string): What is the education level of the students
involved in the study?
- education_setting (string): Provide a brief description of the learning
setting.
- type_edtech (string): Specify what educational technology was used in the
study.
- research_focus (string): Provide a brief summary of the research focus.
```

```
- research_questions (string): List the research questions or study
objectives.
- research_design_general (string): Describe the research design employed in
the study.
- research_methodology_general (string): Describe the research methodology
used.
- research_design_quantitative (string): Provide a description of any
quantitative aspects of the study.
- research_design_qualitative (string): Provide a description of any
qualitative aspects of the study.
- research_design_sample (string): What is the sample size used in the
study?
- research_design_primary_focus_type_string (string): Describe the causal
inference strategies employed.
- research_design_secondary_focus_type_string (string): Describe the
rigorous research methods used.
- comparison_type_string (string): Describe the comparison or control group
characteristics.
- topic_disability_string (string): Specify the types of disabilities or
special educational needs addressed.
- topic_minorities_string (string): Describe the ethnic identities of
students addressed.
- demographic_students_age (string): List the ages of the students studied.
- demographic_students_groups (string): List any specific groups of students
focused on in the study.
- location_country (string): State the country or countries where the study
took place.
- intervention_type_string (string): Summarise any digital approaches used
for interventions.
- learning_outcomes_string (string): List the specific outcome measures
used.
- outcomes_types_string (string): List the standardised tests used to
measure student attainment.
- statistical_significance (string): Present detailed statistical findings
for each outcome mentioned.
- meta-analysis_number (string): If applicable, how many studies were
included in the meta-analysis?
- meta-analysis_effectsize (string): Provide details on effect sizes
reported in the meta-analysis.

Input:
```

```
- Title: {title}
- Abstract: {abstract}

MAKE SURE THE OUTPUT IS A VALID JSON AND HAS NO SYNTAX ERRORS!
```

5.4. Running the Script

The evaluation process was conducted using a Python script that allows for easy configuration and execution across different models. Below is an explanation of how to run the script:

- **Preparation**:
  - Ensure that all necessary libraries, including langchain_together and pandas, are installed.
  - Obtain the API key from Together AI, which is required to access and use the models.
- **Configuration**:
  - Create a JSON settings file that specifies the model configurations, including the model names, temperatures, maximum tokens, and cost per token.
- **Running the Script**:
  - The script can be executed with various command-line arguments to specify the model, prompt, and data source. Here are some examples:

Basic Execution: To run the script using `m1` and `m2` with a predefined prompt:

```
python3 chat_ai.py --settings path/to/settings.json --model m1,m2
--prompt path/to/prompt.txt --data zotero:group_id:collection_id
```

Using a CSV File: To use data from a CSV file:

```
python3 chat_ai.py --settings settings.json --model c1,L2,m1,m2
--prompt prompt.txt --csv data.csv
```

Handling JSON Data: To process a JSON file:

```
python3 chat_ai.py --settings settings.json --model c1,L2,m1,m2
```

```
--prompt prompt.txt --json data.json
```

- **Processing and Output**:
    - The script processes the data by generating prompts based on the provided template and then running them through the specified models.
    - The results are saved in JSON format, with detailed metadata, including the cost of the operation, execution time, and model-specific parameters.

5.5. Evaluation Metrics

The models were evaluated based on several metrics:

- **Accuracy**: The correctness of the extracted information, ensuring that the output JSON is well-structured and free of syntax errors.
- **Cost Efficiency**: The cost per million tokens was calculated for each model, providing insights into the financial implications of using high vs. low parameter settings.
- **Performance**: Execution time was recorded to evaluate the speed of processing across different models and configurations.

5.6. Summary of Findings

The evaluation of these LLMs provided valuable insights into their strengths and weaknesses. LLaMA3, with high parameters, demonstrated exceptional accuracy but at a higher cost. Mixtral, with its mixed-precision approach, offered a good balance between speed and accuracy. ChatGPT 3.5 Turbo stood out for its cost efficiency, making it a viable option for large-scale tasks where cost is a significant concern.

5.7. Tools and other requirements.

The tools and other requirements are provided in the Annex below.

5.8. Conclusion

Finally, the deployment of **Large Language Models (LLMs)** from Together AI demonstrated the potential of AI in extracting and structuring complex information. Each tool and technique was chosen and optimised to meet the specific needs of this research, ensuring that the results are both meaningful and actionable. The careful selection of tools, combined with rigorous testing and validation, has paved the way for a comprehensive and insightful analysis of the research data.

Chapter 6 : Results and Discussion

For this project, the research question is:

How can LDA and LLM contribute to a meta-analysis on EdTech for disadvantaged children?

To answer this research question, the project focused on leveraging Latent Dirichlet Allocation (LDA) and Large Language Models (LLMs) to enhance the meta-analysis process, particularly in the context of evaluating EdTech interventions aimed at improving educational outcomes for disadvantaged children. The use of these advanced AI techniques enabled a more thorough and efficient analysis, providing insights that might not have been achievable through traditional methods alone.

6.1. LDA Contribution

LDA was instrumental in identifying and categorising the thematic structures within the academic literature collected during the systematic review. Here's how LDA contributed to the research:

- **Identification of Core Topics**: LDA facilitated the extraction of key topics related to EdTech interventions, allowing for a structured understanding of the thematic focus areas within the literature. For instance, topics such as "Teacher Professional Development and EdTech" and "Digital Learning Apps" emerged as dominant themes. The intertopic distance maps (as shown in Figure 6.1 and Figure 6.2) illustrate the relationships and distances between different topics, highlighting the coherence and distinctiveness of each theme.

Figure 6.1. *Intertopic Distance Map and Top-30 Most Relevant Terms for Topic 5*

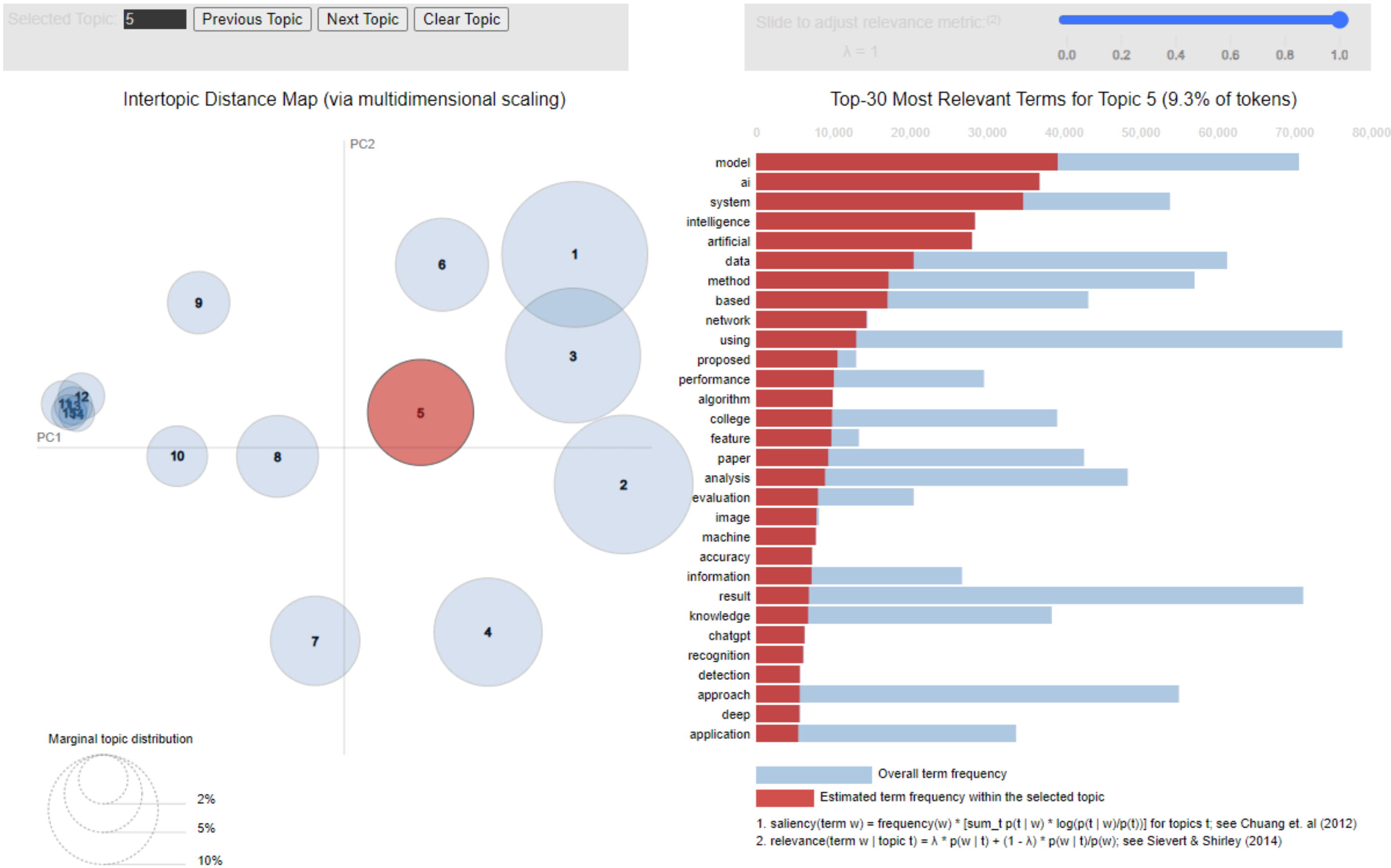
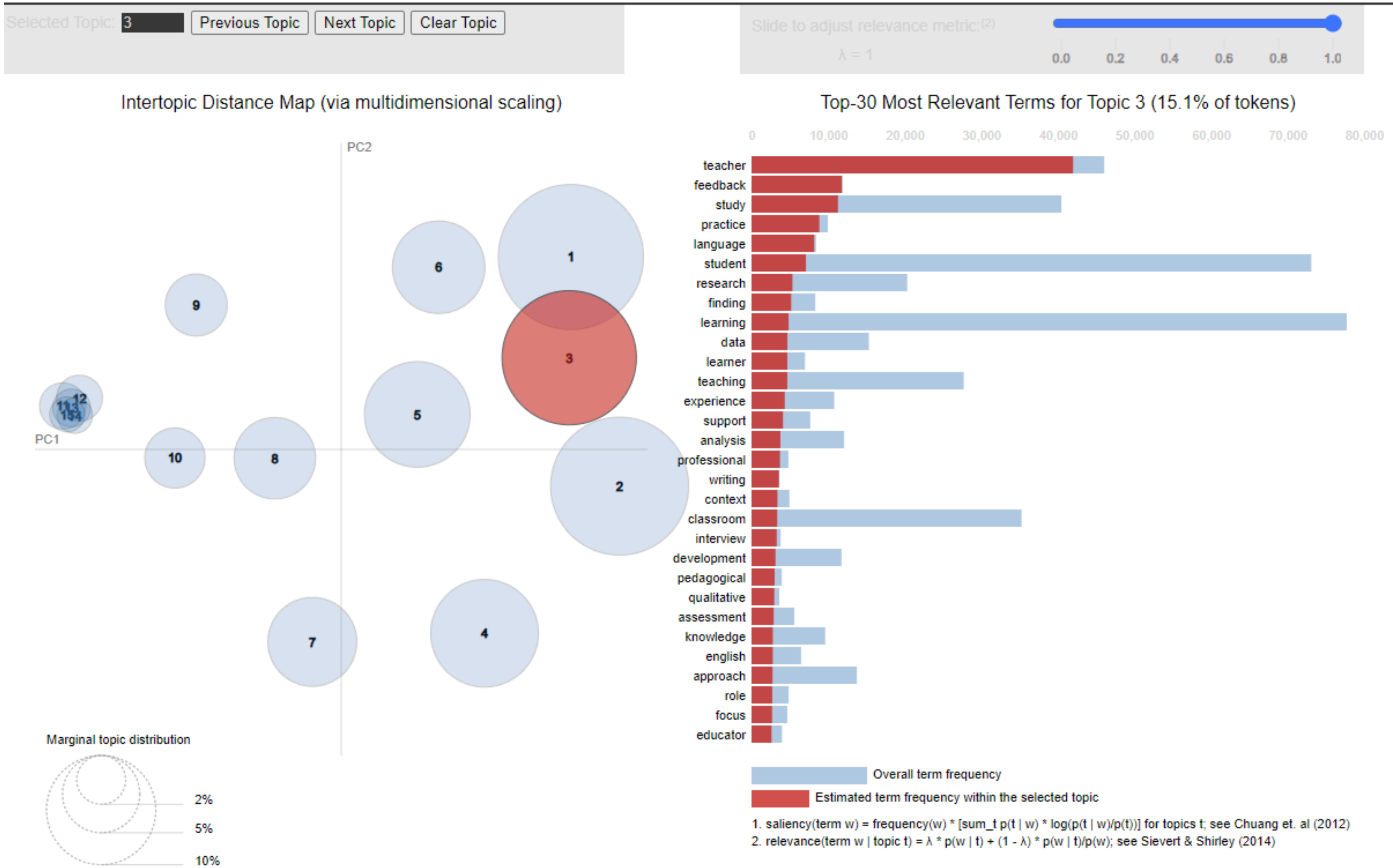


Figure 6.2. *Intertopic Distance Map and Top-30 Most Relevant Terms for Topic 3*



- **Topic Validation and Refinement**: Through iterative refinement, LDA ensured that the topics identified were both relevant and aligned with the research objectives. This process added an extra layer of rigour to the traditional systematic review approach, confirming the pertinence of the selected studies.


```json
{
  "Paper 43416": {
    "Title": "Efficacy of video-based teacher professional development for increasing classroom discourse and student learning",
    "Topics": {
      "Hybrid learning models": 0.118633464,
      "Teacher professional development and EdTech": 0.42114663,
      "Digital learning apps": 0.272325
    },
    "Top Validated Topics": {
      "Teacher professional development and EdTech": 0.42114663,
      "Digital learning apps": 0.272325,
      "Hybrid learning models": 0.118633464
    }
  }
}

{
  "Paper 43417": {
    "Title": "Open educational resources for CALL teacher education: the iTILT interactive whiteboard project",
    "Topics": {
      "Game-based learning": 0.31701797,
      "Teacher professional development and EdTech": 0.3582981,
      "Curriculum Enhancement through EdTech": 0.21681602
    },
    "Top Validated Topics": {
      "Teacher professional development and EdTech": 0.3582981,
      "Game-based learning": 0.31701797,
      "Curriculum Enhancement through EdTech": 0.21681602
    }
  }
}
```

```
{
  "Paper 43418": {
    "Title": "Influence of gender and computer teaching efficacy
on computer acceptance among Malaysian student teachers: An
extended technology acceptance model",
    "Topics": {
      "Game-based learning": 0.11583611,
      "Hybrid learning models": 0.6058227
    },
    "Top Validated Topics": {
      "Hybrid learning models": 0.6058227,
      "Game-based learning": 0.11583611
    }
  }
}

{
  "Paper 43419": {
    "Title": "Investigating cognitive holding power and equity in
the flipped classroom",
    "Topics": {
      "Game-based learning": 0.107142255,
      "Teacher professional development and EdTech": 0.19049487,
      "Digital learning apps": 0.5441655
    },
    "Top Validated Topics": {
      "Digital learning apps": 0.5441655,
      "Teacher professional development and EdTech": 0.19049487,
      "Game-based learning": 0.107142255
    }
  }
}
```

These examples demonstrate how LDA's iterative refinement process added an extra layer of rigour to the traditional systematic review approach, confirming the pertinence of the selected studies and ensuring that the analysis remained closely aligned with the research objectives.

- **Creation of Evidence Maps**: The topics generated by LDA were visualised through evidence maps, providing a clear overview of the research landscape. This visual representation helped in identifying research gaps and areas with sufficient data for further meta-analysis.

For instance, the evidence maps (as shown in **Figure 5.3** and **Figure 5.4**) depict the distribution of various EdTech topics across numerous studies, highlighting which areas are well-researched and which might require further investigation.

Figure 6.3. *Evidence Map - Topic Distribution Across Papers*

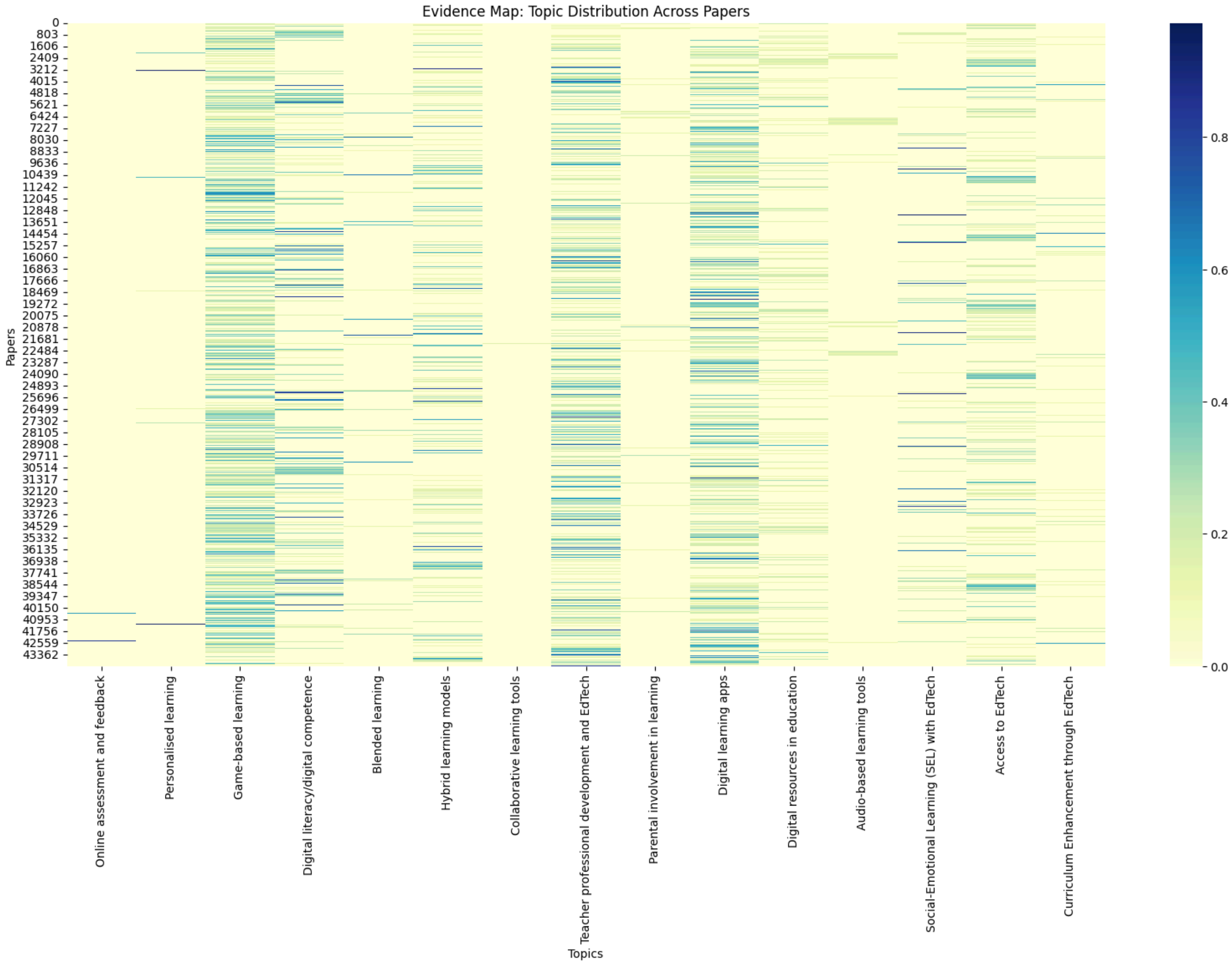
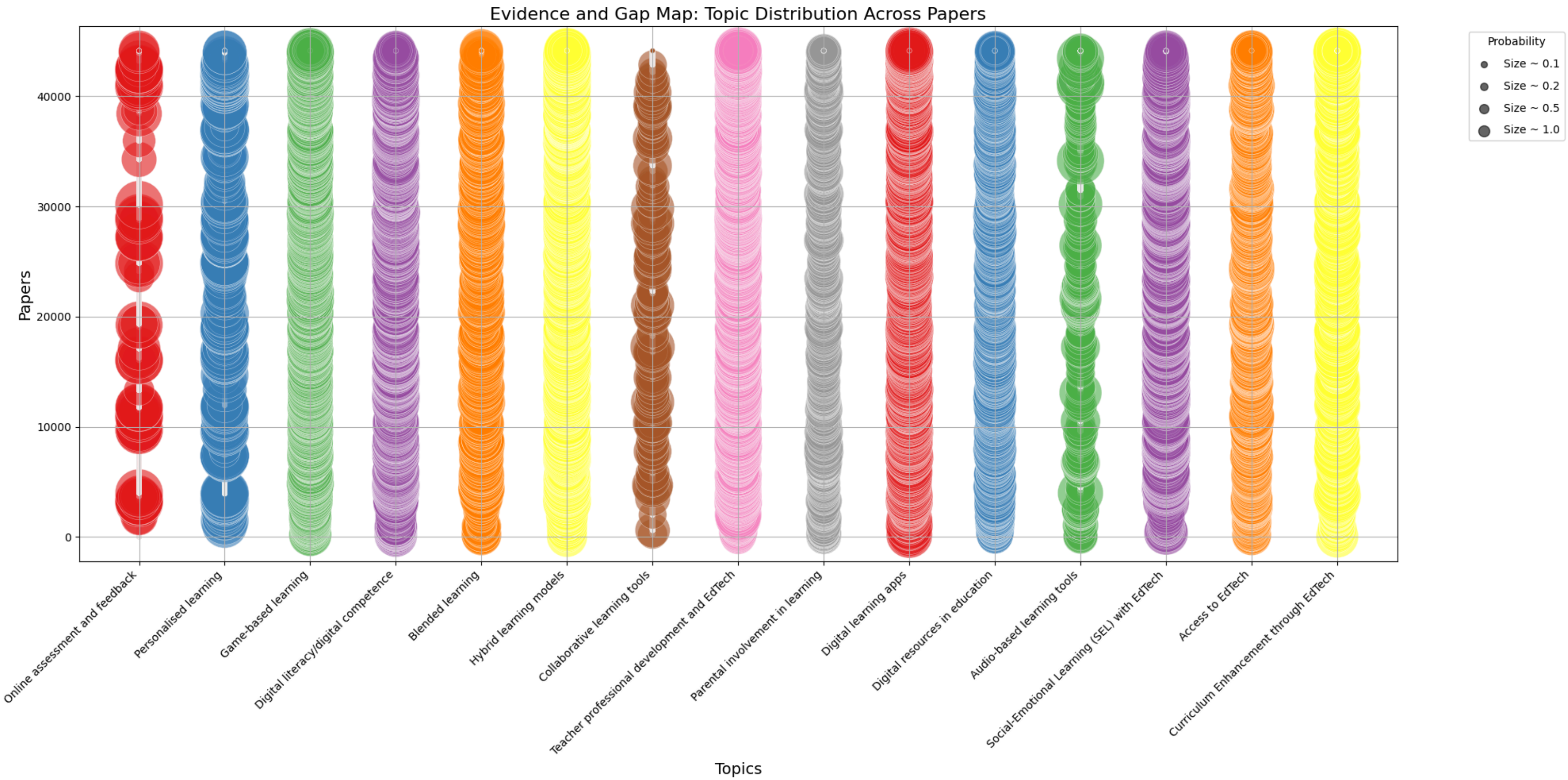


Figure 6.4. *Evidence and Gap Map - Topic Distribution Across Papers*



The following are some of the topics that were identified as having sufficient studies for further analysis:

```
{
  "Topics with Sufficient Studies for Further Analysis": {
    "Online assessment and feedback": 61,
    "Personalised learning": 299,
    "Game-based learning": 15074,
    "Digital literacy/digital competence": 6197,
    "Blended learning": 911,
    "Hybrid learning models": 5417,
    "Collaborative learning tools": 60,
    "Teacher professional development and EdTech": 10222,
    "Parental involvement in learning": 118,
    "Digital learning apps": 10558,
    "Digital resources in education": 2558,
    "Audio-based learning tools": 228,
    "Social-Emotional Learning (SEL) with EdTech": 2486,
    "Access to EdTech": 4928,
    "Curriculum Enhancement through EdTech": 668
  }
}
```

These evidence maps were instrumental in identifying where there is a concentration of research and where further studies could be beneficial. For example, "Game-based learning" and "Teacher professional development and EdTech" had a substantial number of studies, making them well-suited for deeper meta-analytical exploration. In contrast, topics such as "Collaborative learning tools" and "Audio-based learning tools" had fewer studies, indicating potential areas for further research and exploration.

6.2. LLM Contribution

LLMs, particularly those from Together AI, played a critical role in automating the screening and extraction processes, which are traditionally time-consuming tasks in systematic reviews. The use of LLMs contributed to the study in several ways:

- **Automated Screening**: LLMs were used to automatically screen titles, abstracts, and keywords, significantly reducing the manual workload. The models efficiently filtered out irrelevant studies and prioritised those that met the inclusion criteria, as demonstrated in the results extracted from the LLM, formatted as JSON.
- **Precision in Screening**: By employing instruct models, the LLMs were fine-tuned to follow specific prompts and screening questions, ensuring that the selected studies were highly relevant to the research question. This led to more consistent and accurate screening results compared to manual processes.
- **Detailed Information Extraction**: LLMs extracted detailed information from the studies, such as research approaches, topics, demographic focus, and intervention types. For example, one of the JSON outputs indicated that the studies predominantly employed quantitative methodologies, focused on EdTech and literacy for young children, and used systematic reviews and meta-analyses to assess the impact of electronic storybooks.


```
{
    "topic_language": true,
    "research_approach_quantitative": true,
    "research_approach_rct": false,
    "research_approach_qed": true,
    "research_approach_impact_evaluation": true,
    "research_approach_process_evaluation": false,
    "research_approach_qualitative": false,
    "research_approach_mixed_methods": false,
    "research_approach_literature_review": false,
    "research_approach_systematic_review": true,
    "research_approach_meta_analysis": true,
    "research_design_primary_focus_type": true,
    "topic_education": true,
    "topic_edtech": true,
    "topic_disadvantaged": false,
    "demographic_preschool_students": true,
    "geographic_focus_techready": true,
    "intervention_type": true,
    "subject_lit": true,
    "learning_outcomes": true,
    "Outcomes_types": true,
    "topic_subject": ["Literacy"],
```

```
    "education_level": "Pre-K to grade 2",
    "education_setting": "Electronic storybooks",
    "type_edtech": "Electronic storybooks",
    "research_focus": "Effects of electronic storybooks on
language and literacy outcomes for children in grades Pre-K to
grade 2",
    "meta-analysis_effectsize": "Effect sizes reported in the
meta-analysis"
}
```

- **Enhanced Data Extraction**: LLMs facilitated the extraction of structured data from the selected studies, allowing for a more systematic and detailed analysis. The ability of LLMs to follow complex instructions and produce outputs in structured formats (e.g., JSON) streamlined the data extraction process.
- **Iterative Improvement**: The AI-driven screening and extraction processes were iteratively refined through feedback loops between human reviewers and the LLMs. This iterative approach enhanced the reliability of the results, ensuring that the final selection of studies was both comprehensive and pertinent.

6.3. Implications for Disadvantaged Students

The identified topics indicate that teacher professional development, digital learning apps, and game-based learning are critical components of successful EdTech interventions. These tools are particularly effective in supporting teachers and students in disadvantaged settings by:

- **Enhancing Teacher Capacity**: By improving teacher training and support, EdTech can bridge the gap between traditional teaching methods and modern educational demands, ensuring that even in disadvantaged areas, teachers are equipped with the necessary skills to leverage technology effectively.
- **Promoting Equity**: The use of digital learning apps and hybrid learning models helps level the playing field for students from different backgrounds, providing them with equal opportunities to access quality education.
- **Engaging Learning Environments**: Game-based learning, in particular, is shown to enhance student engagement and motivation, which is crucial for

learners in disadvantaged settings where traditional methods may not be as effective.

6.4. Implications for Future Research

The comparative analysis of AI and human screening presents significant implications for the future of systematic reviews and meta-analyses. The findings suggest that while LLMs offer substantial benefits in terms of efficiency and consistency, they are most effective when used in conjunction with human expertise. This hybrid approach could become a standard in future research, where AI tools are employed to handle large-scale data processing tasks, freeing human researchers to focus on more complex and judgement-intensive aspects of the review process. Further research could explore how to optimise this collaboration, potentially developing new AI models that better replicate human decision-making or more sophisticated ways to integrate AI recommendations into the human review workflow.

Chapter 7. Conclusion

This study set out to explore the potential of integrating Latent Dirichlet Allocation (LDA) and Large Language Models (LLMs) in the systematic review and meta-analysis of EdTech interventions, specifically focusing on their impact on disadvantaged children.

The integration of LDA and LLMs into the meta-analysis process has demonstrated the potential of AI to transform traditional research methodologies. The use of these technologies not only enhanced the efficiency of the systematic review but also improved the depth and accuracy of the analysis. Specifically, LDA provided a robust framework for topic modelling, ensuring that the literature was thoroughly categorised and validated. Meanwhile, LLMs automated key aspects of the review process, reducing the time and resources required for manual screening and data extraction.

The results of this study indicate that LDA and LLMs can significantly contribute to the meta-analysis of EdTech interventions, particularly in the context of disadvantaged children. By enabling a more rigorous and efficient review process, these technologies help ensure that the conclusions drawn from the meta-analysis are based on a comprehensive and accurate synthesis of the available evidence. This, in turn, supports the development of evidence-based recommendations for the implementation of EdTech in educational settings, particularly those aimed at improving outcomes for disadvantaged students.

Applying LDA and LLMs in this study highlights AI's transformative potential in educational research. These technologies not only streamline the review process but also enhance the quality and relevance of the findings, providing valuable insights into the mechanisms that drive the success of EdTech interventions.

Key Findings and Contributions

The application of LDA allowed for the identification and refinement of relevant topics within the vast corpus of academic literature on EdTech interventions. The iterative validation process ensured that the topics extracted were not only statistically significant but also aligned with the research objectives. This approach provided a robust foundation for creating an evidence map that visually represented the research landscape, highlighting areas of sufficient data for further analysis and identifying critical gaps.

LLMs, particularly those available at Together AI, were instrumental in streamlining the screening and extraction processes. By leveraging these advanced models, the study was able to efficiently filter and categorise a large number of studies, ensuring that only the most relevant and high-quality research was included in the final analysis. This AI-driven approach significantly reduced the manual workload and improved the consistency and accuracy of the review process.

Given that there is only one published paper (↑Spillias et al., 2023) and one preprint (↑Nakaya et al., 2023), this thesis significantly extends the use of LLMs in meta-analysis. The paper by ↑Spillias et al. (2023) explored the application of LLMs in systematic reviews, focusing on human-AI collaboration to improve the screening and selection of literature. They found that using multiple AI replicates and reflection techniques enhanced the accuracy and consistency of the review process. In my thesis, I extended this approach by applying LLMs specifically to the systematic review and meta-analysis of EdTech interventions aimed at disadvantaged children. My thesis not only utilised LLMs for screening but also employed them for structured data extraction and synthesis, demonstrating their utility in a comprehensive meta-analysis framework.

In the preprint by (↑Nakaya et al. (2023), the authors tested ChatGPT's ability to classify virtual reality studies in cardiology, comparing its performance with traditional NLP methods. They found that while LLMs like ChatGPT were highly accurate, there were challenges related to the interpretability of the AI's decision-making process. In my thesis, I addressed similar challenges by refining the prompts and validation processes used with LLMs to ensure that the extracted data was both accurate and interpretable. Furthermore, I integrated these AI-driven classifications with LDA-based topic modelling, offering a more holistic approach to analysing the impact of EdTech interventions on disadvantaged students.

Limitations and Future Research

While the integration of LDA and LLMs in this study yielded significant benefits, there are limitations to consider. The quality of the output is heavily dependent on the quality of the input data. Biases inherent in the training data of LLMs could potentially influence the results. Additionally, while LDA is effective for topic modelling, it may struggle with interpretability and the need for extensive parameter tuning.

Future research could explore the use of alternative topic modelling techniques or hybrid approaches that combine the strengths of different models. Further investigation into the ethical considerations surrounding the use of AI in educational research is also warranted, particularly in terms of data privacy and bias mitigation.

Conclusion

In conclusion, this study demonstrates the potential of combining traditional systematic review methods with cutting-edge AI technologies to advance educational research and practice. The integration of LDA and LLMs provided a powerful framework for analysing and synthesising the existing literature on EdTech interventions, offering valuable insights into the mechanisms that drive educational success. As the field of EdTech continues to evolve, the methodologies developed in this study will serve as a foundation for future meta-analysis, therefore making important contributions to research analysis, such as the use of educational technologies.

References

This bibliography is available digitally in our evidence library at https://docs.opendeved.net/lib/BL6WUMNW

Bulman, G., & Fairlie, R. W. (2016). Chapter 5 - Technology and Education: Computers, Software, and the Internet. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the Economics of Education* (Vol. 5, pp. 239–280). Elsevier. https://doi.org/10.1016/B978-0-444-63459-7.00005-1. Available from https://www.sciencedirect.com/science/article/pii/B9780444634597000051. (details)

Escueta, M., Quan, V., Nickow, A. J., & Oreopoulos, P. (2017). *Education technology: An evidence-based review*. https://www.nber.org/papers/w23744. (details)

Haßler, B., Adam, T., Allier-Gagneur, Z., Blower, T., Brugha, M., Damani, K., Hennessy, S., Hollow, D., Jordan, K., Martin, K., Megha-Bongnkar, G., Murphy, M., & Walker, H. (2021k). *Methodology for literature reviews* (Working Paper No. 10). EdTech Hub. https://doi.org/10.5281/zenodo.4557508. Available from https://docs.edtechhub.org/lib/2CKWI7RR. Available under Creative Commons Attribution 4.0 International. (details)

Haßler, B., Hassan, M., & Klune, C. (2024). *Using AI to Automate Literature Reviews* [Working paper]. EdTech Hub. https://doi.org/10.53832/edtechhub.1003. Available from https://docs.edtechhub.org/lib/BVD8JX7V. Available under Creative Commons Attribution 4.0 International. (details)

Haßler, B., McBurnie, C., Walker, H., Klune, C., Huntington, B., & Bhutoria, A. (2024). *Protocol for a systematic review with meta-analysis: Understanding quality characteristics of EdTech interventions and implementation for disadvantaged pupils* (Understanding Quality Characteristics of EdTech Interventions and Implementation for Disadvantaged Pupils No. 1). Open Development & Education. https://doi.org/10.53832/opendeved.1077. Available from https://docs.opendeved.net/lib/2I2GT22T. (details)

ISTE. (2020). 2. Educators. In *ISTE*. https://iste.org/standards/educators. (details)

Nakaya, Y., Higaki, A., & Yamaguchi, O. (2023). ChatGPT's ability to classify virtual reality studies in cardiology. *European Heart Journal - Digital Health*, *4*(3), 141–142. https://doi.org/10.1093/ehjdh/ztad026. (details)

Nussbaumer-Streit, B., Ellen, M., Klerings, I., Sfetcu, R., Riva, N., Mahmić-Kaknjo, M., Poulentzas, G., Martinez, P., Baladia, E., Ziganshina, L. E., Marqués, M. E., Aguilar, L., Kassianos, A. P., Frampton, G., Silva, A. G., Affengruber, L., Spjker, R., Thomas, J., Berg, R. C., … Gartlehner, G. (2021). Resource use during systematic review production varies widely: a scoping review. *Journal of Clinical Epidemiology*, *139*, 287–296. https://doi.org/10.1016/j.jclinepi.2021.05.019. Available from https://linkinghub.elsevier.com/retrieve/pii/S0895435621001712. (details)

OECD. (2021). *Education at a Glance 2021: OECD Indicators*. Organisation for Economic Co-operation and Development. https://www.oecd-ilibrary.org/education/education-at-a-glance-2021_b35a14e5-en. (details)

Spillias, S., Tuohy, P., Andreotta, M., Annand-Jones, R., Boschetti, F., Cvitanovic, C., Duggan, J., Fulton, E., Karcher, D., & Paris, C. (2023). *Human-AI Collaboration to Identify Literature for Evidence Synthesis*. https://doi.org/10.21203/rs.3.rs-3099291/v1. Available from https://www.researchsquare.com/article/rs-3099291/latest. (details)

Vanbecelaere, S., Adam, T., Sieber, C., Clark-Wilson, A., Boody Adorno, K., & Haßler, B. (2023). *Towards Systemic EdTech Testbeds: A Global Perspective*. Global EdTech Testbeds Network. https://doi.org/10.53832/opendeved.0285. Available from https://docs.opendeved.net/lib/78XN44TG. Available under Creative Commons Attribution 4.0 International (https://creativecommons.org/licenses/by/4.0/). (details)

Wang, H., Jia, S., Li, Z., Duan, Y., Tao, G., & Zhao, Z. (2022). A Comprehensive Review of Artificial Intelligence in Prevention and Treatment of COVID-19 Pandemic. *Frontiers in Genetics*, *13*, 845305. https://doi.org/10.3389/fgene.2022.845305. Available from https://www.frontiersin.org/articles/10.3389/fgene.2022.845305/full. (details)

Klune, C., Huntington, B., Touiyate, A., … Haßer, B. - 2024 - Open Development & Education, Education Endowment Foundation. 10.53832/opendeved.1132 Available from https://docs.opendeved.net/lib/P527M7BW

Annex: Tools and Requirements

To successfully implement the methods and processes described in this chapter, a combination of software tools and specific system requirements are necessary. Below is a list of the key tools and their respective roles:

Tools

Description: Python 3.x is the primary programming language used for implementing scripts and models in this research.

Figure 9.1. *Python 3.x*

Description: VS Code is a versatile code editor used for writing and testing the scripts.

Figure 9.2. *VS Code (Visual Studio Code)*

Description: Google Colab provides free access to GPUs and TPUs for running and testing models.

Figure 9.3. *Google Colab*

Description: Pandas is a data manipulation and analysis library in Python used for handling and processing CSV and JSON files.

Figure 9.4. *Pandas*

Description: JQ is a command-line tool used for processing and filtering JSON data.

Figure 9.5. *JQ (Command-line Tool)*

Figure 9.6. *Together AI*

Description: Together AI provides access to various LLMs for natural language processing tasks.

Discripsion: LangChain_Together is a Python library used to interface with LLMs on Together AI.

Figure 9.7. *LangChain_Together*

Description: Scholarly and ProxyGenerator are Python libraries used in the Scholarly-cli tool, which interacts with Google Scholar to retrieve academic papers and related metadata.

Figure 9.8. *Scholarly and ProxyGenerator*

Description: Zotero is a tool used for managing and sharing research references, and it is utilised in the script to fetch and organise data from research collections. The bibliography was compiled using bZotBib.

Figure 9.9. *Zotero*

Description: OpenAlex is an index of scholarly papers, authors, and institutions. It is used in the script to fetch academic data for processing.

Figure 9.10. *OpenAlex*

Installation of Dependencies

Dependencies can be installed using the following command:

```
pip install langchain_together pandas jq scholarly
```

Ensure all dependencies are properly installed and configured before running the scripts.

System Requirements

- **Operating System**: Compatible with Windows, macOS, or Linux.
- **Memory**: Adequate memory is necessary, particularly when running models with high parameters or processing large datasets.
- **Network Connection**: A stable internet connection is required to interact with the Together AI API, Google Scholar, and external data sources.

Abstract

This report explores the integration of artificial intelligence (AI) into systematic reviews and meta-analyses, specifically focusing on educational technology (EdTech) interventions aimed at improving student outcomes. The research combines traditional evidence synthesis methods with advanced AI tools, such as Latent Dirichlet Allocation (LDA) for topic modeling and Large Language Models (LLMs) for automated literature screening. The goal is to enhance the efficiency and accuracy of the review process, ultimately providing a structured overview of successful EdTech interventions. The findings offer practical recommendations for educators, policymakers, and researchers in the field of education.

Résumé

Ce rapport explore l'intégration de l'intelligence artificielle (IA) dans les revues systématiques et les méta-analyses, en se concentrant spécifiquement sur les interventions en technologies éducatives (EdTech) visant à améliorer les résultats des élèves. La recherche combine les méthodes traditionnelles de synthèse des preuves avec des outils d'IA avancés, tels que l'Allocation Latente de Dirichlet (LDA) pour la modélisation des sujets et les Modèles de Langage à Grande Échelle (LLM) pour le criblage automatisé de la littérature. L'objectif est d'améliorer l'efficacité et la précision du processus de revue, en offrant finalement une vue d'ensemble structurée des interventions EdTech réussies. Les résultats offrent des recommandations pratiques pour les éducateurs, les décideurs politiques et les chercheurs dans le domaine de l'éducation.