| **Section and Topic** | **Item #** | **Checklist item** | **Location where item is reported** |
| --- | --- | --- | --- |
| **TITLE** | | |  |
| Title | 1 | Natural Language Processing with Transformer-Based Models: A Meta-Analysis | Page 1 |
| **ABSTRACT** | | |  |
| Abstract | 2 | Included a summary of the background, objectives, methods, results, and conclusion of the study. | Page 1 |
| **INTRODUCTION** | | |  |
| Rationale | 3 | The aim of this paper was to analyze the performance of transformer-based models across various NLP tasks, their scalability, domain adaptation, and the ethical implications of such models. | Page 1 |
| Objectives | 4 | The research questions were as follows:   1. How do transformer-based models perform across various natural language processing tasks? 2. What are the scalability and domain adaptation capabilities of transformer-based models in NLP applications? 3. What are the key challenges and ethical considerations in developing and applying transformer-based models for NLP tasks? | Page 4 |
| **METHODS** | | |  |
| Eligibility criteria | 5 | Inclusion criteria included (1) studies that involve transformer-based architectures, (2) studies reporting empirical results on NLP tasks using transformers, and (3) studies providing sufficient methodological details for replication or meta-analysis. Exclusion criteria included (1) studies focusing solely on theoretical aspects without empirical validation, (2) studies involving outdated architectures or models, and (3) non-English publications or studies with incomplete data. | Page 5 |
| Information sources | 6 | The databases were: Springer, Elsevier, PubMed, IEEE Xplore, ACM Digital Library, and Google Scholar. | Page 5 |
| Search strategy | 7 | The literature search was conducted systematically across multiple electronic databases, including Springer, Elsevier, PubMed, IEEE Xplore, ACM Digital Library, and Google Scholar. Search terms were selected to encompass various aspects of Transformer research, such as “Transformer NLP,” “Transformer scalability,” “domain adaptation in Transformers,” and “ethical considerations in NLP models.” Boolean operators were employed to refine the queries and maximize the retrieval of relevant articles. The search was supplemented by manually screening the reference lists of included articles to identify additional studies. Grey literature, including technical reports and unpublished dissertations, was also reviewed to capture less formal but potentially valuable contributions | Page 5 |
| Selection process | 8 | Study selection followed a two-stage process. In the first stage, titles and abstracts were screened against the inclusion criteria to exclude irrelevant studies. This process was conducted systematically to reduce the risk of bias, with careful attention to the predefined inclusion criteria. Any uncertainties or ambiguous cases were resolved through further scrutiny of the full text and by cross-referencing related studies to ensure alignment with the scope of the meta-analysis. The second stage involved a full-text review of the remaining articles to ensure they met all eligibility criteria. A standardized form was used to document decisions at each stage, and the reasons for excluding studies were recorded to enhance transparency | Page 6 |
| Data collection process | 9 | Data extraction was done using a pre-designed template to ensure consistency and standardization across all selected studies. The template was structured to capture key details such as model architecture, dataset characteristics, evaluation metrics, task-specific performance, computational costs, and reported challenges. The most relevant and representative results were prioritized for studies presenting multiple experiments or configurations, focusing on those aligned closely with the research objectives. | Page 7 |
| Study risk of bias assessment | 10 | Publication bias was assessed using funnel plots and Egger’s test. This step ensured that the results were not unduly influenced by selective reporting or other forms of bias. | Page 7 |
| Synthesis methods | 11a | A mixed-methods approach was adopted for data synthesis, combining quantitative and qualitative techniques. | Page 7 |
| 11b | Quantitative data were synthesized through meta-analytic techniques, calculating pooled effect sizes and confidence intervals where sufficient homogeneity existed. | Page 7 |
| 11c | Qualitative data were synthesized using thematic analysis, identifying recurring patterns and themes across studies. | Page 7 |
| 11d | Heterogeneity was assessed using the I² statistic. | Page 7 |
| 11e | Subgroup analyses were conducted to explore variations across different model types, tasks, and domains. | Page 7 |
| 11f | The findings were reported with transparency and precision, avoiding overgeneralizations or unwarranted extrapolations. | Page 7 |
| Certainty assessment | 12 | The quality of the included studies was assessed using the modified version of the Newcastle-Ottawa Scale, adapted for computational research, Studies scoring below a predefined threshold were excluded from the analysis to maintain the integrity of the findings. | Page 7 |
| RESULTS | | |  |
| Study selection | 13 | The review included 25 studies selected from 136 records, as documented in a PRISMA flowchart. Reasons for exclusion included outdated architectures, lack of empirical results, or incomplete data. | Page 6 |
|  |  |  |
| Study characteristics | 14 | Studies that met the inclusion criteria were included in the study and characteristics of each included study were tabulated by task and year. | Page 15 -16 |
| Risk of bias in studies | 15 | Risk of bias was assessed using a modified Newcastle-Ottawa Scale. | Page 7 |
| Results of individual studies | 16 | For each NLP task—such as translation, summarization, QA, sentiment analysis, NER, and generation—summary statistics and effect estimates (e.g., BLEU, ROUGE, F1) were presented. Tables detailed performance and limitations (e.g., hallucination in summarization, domain shift in NER). | Page 8 |
| Results of syntheses | 17a | Meta-analyses showed transformers outperform older models, though with high variance in resource demands and generalization. | Page 8 |
| 17b | Heterogeneity causes were explored through subgroup analysis by task and model type. | Page 9 |
| 17c | Sensitivity analyses included comparisons of pretraining strategies and fine-tuning methods | Page 10 |
| Reporting biases | 18 | Reporting bias was assessed using funnel plots. | Page 11 |
| Certainty of evidence | 19 | Evidence certainty was discussed based on dataset coverage, task alignment, and reproducibility. | Page 12 |
| **DISCUSSION** | | |  |
| Discussion | 20 | Transformer models consistently outperform traditional approaches across key NLP tasks, aligning with existing literature. | Page 13 |
| 21 | However, issues like hallucination, bias, and high resource demands limit reliability and inclusivity. | Page 14 |
| 22 | The review's scope may be constrained by publication bias and language exclusions. | Page 14 |
| 23 | These findings highlight the need for ethical safeguards, domain-specific tuning, and research into scalable, fair, and efficient architectures to guide future practice and policy. | Page 14 |
| **OTHER INFORMATION** | | |  |
| Support | 24 | The author(s) received no specific funding for this study. | Page 14 |
| Competing interests | 25 | The authors declare they have no conflicts of interest to report regarding the present study | Page 15 |

*From:*  Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi: 10.1136/bmj.n71. This work is licensed under CC BY 4.0. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>