SCOPUS Query Documentation

We began the Block A search with the goal of capturing as much relevant academic work as possible on open-source software and AI. To achieve this, we intentionally began with a broad query that encompassed general open-source terminology, as well as common AI and machine learning terms, such as “machine learning,” “deep learning,” and “neural networks.” This approach is standard in systematic reviews: it ensures that we do not miss important material at the outset, even if it means casting the net a bit wider than initially needed.

However, when this first version was run in Scopus, it returned around 100,000 documents. It quickly became clear that many of these hits were from fields such as medicine, biology, environmental science, and remote sensing, where terms like “open source” or “deep learning” frequently appear but have no connection to software security. While the initial breadth confirmed good recall, the scale and noise were too high to work with efficiently and would not support a transparent or manageable screening process.

To address this, we refined the query step by step. Our first adjustment was to add a “software context” layer, terms like “software,” “system*,” “platform*,” and “code.” This was designed to keep the search within the technical and software-development landscape. Although sensible in principle, this change only slightly reduced the result count. We realised that Scopus was still picking up a wide range of unrelated material, meaning the search terms were still too general and the Boolean logic wasn’t being enforced as intended.

The breakthrough came when we corrected the structure of the query itself. Scopus only supports a single TITLE-ABS-KEY() field per search block. Earlier tests with multiple declarations caused Scopus to interpret parts of the query separately, resulting in broader retrieval than we intended. Once we placed all conceptual elements, open-source terms, software-context terms, and AI terms inside the same TITLE-ABS-KEY() block, Scopus correctly applied AND logic across the entire query. This dramatically reduced the results, but it also revealed that one version of our AI block (“open source AI,” “open LLM,” etc.) was too narrow, producing only about 275 results. It became clear that this newer terminology is not yet widely used across the academic literature.

We therefore broadened the AI component again, but in a more targeted way. Instead of reintroducing all general AI terms (which we knew would overwhelm the search again), we focused on terminology associated with specific model architectures, phrases like “AI model*,” “ML model*,” “foundation model*,” “large language model*,” and “LLM.” This brought the count back up to roughly 17,000, a significant improvement but still on the higher side.

The final refinement involved removing two high-noise terms, “deep learning” and “neural network*”, which were responsible for pulling in thousands of domain-irrelevant papers. After removing these, the result set stabilised in the region of 10,000 to 12,000 documents. This aligns well with the expected size for Block A: large enough to ensure coverage across OSS and AI technologies, but not so large that it becomes operationally unmanageable or undermines later screening stages.

Throughout this process, we deliberately avoided methods that would compromise the integrity of the review. For example, we did not filter by Scopus subject area because subject areas are inconsistently assigned and would have risked excluding governance, policy, and security-related work that is central to DSIT’s research questions. We also avoided restricting searches to titles only, which drastically reduces recall, and we did not prematurely add security-related terms, as these properly belong in Blocks B-E. Likewise, removing AI entirely was never an option, as open-source AI and model-related security risks are explicitly part of the project scope.

The final Block A query strikes a careful balance: it captures the full breadth of open-source and AI technologies while remaining focused on the software context, uses syntax that Scopus interprets correctly, and avoids the noise introduced by overly generic AI terms. It is transparent, reproducible, and methodologically robust, providing a solid foundation for combining with the remaining blocks in the overall WP1 search strategy.

("open source" OR "open-source" OR OSS OR FOSS OR "free software" OR "open source software" OR "open source development" OR "open source component*" OR "open source librar*" OR "open source ecosystem*" OR "open source project*" AND "artificial intelligence" OR AI OR "machine learning" OR ML OR "deep learning" OR "neural network*" OR "large language model*" OR LLM OR "generative AI" OR "foundation model*" OR "pre-trained model*" OR "AI system*" OR "AI model*" OR "open source AI" OR "open-source AI") AND (software OR system* OR platform* OR model OR component*) AND PUBYEAR > 2019 AND PUBYEAR < 2026 AND (LIMIT-TO (EXACTKEYWORD , "Deep Learning") OR LIMIT-TO (EXACTKEYWORD , "Machine Learning") OR LIMIT-TO (EXACTKEYWORD , "Machine-learning") OR LIMIT-TO (EXACTKEYWORD , "Learning Systems") OR LIMIT-TO (EXACTKEYWORD , "Artificial Intelligence") OR LIMIT-TO (EXACTKEYWORD , "Open Source Software") OR LIMIT-TO (EXACTKEYWORD , "Open Systems") OR LIMIT-TO (EXACTKEYWORD , "Open-source") OR LIMIT-TO (EXACTKEYWORD , "Learning Algorithms") OR LIMIT-TO (EXACTKEYWORD , "Neural-networks") OR LIMIT-TO (EXACTKEYWORD , "Software") OR LIMIT-TO (EXACTKEYWORD , "Neural Networks") OR LIMIT-TO (EXACTKEYWORD , "Artificial Neural Network") OR LIMIT-TO (EXACTKEYWORD , "Reinforcement Learning") OR LIMIT-TO (

EXACTKEYWORD , "Natural Language Processing") OR LIMIT-TO (EXACTKEYWORD , "Support Vector Machine") OR LIMIT-TO (EXACTKEYWORD , "Support Vector Machines") OR LIMIT-TO (EXACTKEYWORD , "Machine Learning Models"))

Number of shown records: 106,888 records

TITLE-ABS-KEY ("open source" OR "open-source" OR OSS OR FOSS OR "free software" OR "open source software" OR "open source development" OR "open source component*" OR "open source librar*" OR "open source ecosystem*" OR "open source project*") AND PUBYEAR > 2019 AND PUBYEAR < 2026

Number of shown records: 87,657 records

TITLE-ABS-KEY (("open source" OR "open-source" OR OSS OR FOSS OR "free software" OR "open source software" OR "open source development") AND (software OR system* OR platform* OR code) AND ("artificial intelligence" OR "machine learning" OR "AI model*" OR "ML model*" OR "deep learning" OR "neural network*" OR "foundation model*" OR "large language model*" OR LLM)) AND PUBYEAR > 2019 AND PUBYEAR < 2026

Number of shown records: 17,021 records

TITLE-ABS-KEY (("open source" OR "open-source" OR OSS OR FOSS OR "free software" OR "open source software" OR "open source development") AND (software OR system* OR platform* OR code) AND ("artificial intelligence" OR "machine learning" OR "AI model*" OR "ML model*" OR "large language model*" OR LLM OR "foundation model*")) AND PUBYEAR > 2019 AND PUBYEAR < 2026

Number of shown records: 12,543 records

```
TITLE-ABS-KEY ( ( "open source" OR "open-source" OR OSS OR FOSS OR "free software" OR "open source software" OR "open source development" ) AND ( software OR system* OR platform* OR code ) AND ( "AI model*" OR "ML model*" OR "large language model*" OR LLM OR "foundation model*" ) ) AND PUBYEAR > 2019 AND PUBYEAR < 2026
```

Number of shown records: 3,378 records