When evaluating tools for supporting requirements engineering tasks, we should keep the following four points in mind:

- Do not reinvent the wheel: tool evaluation is not a requirements engineering specific problem. There is plenty of work about "evaluating the evaluation" in other communities such as Natural Language Processing, Information Retrieval and Recommender Systems. RE community should check and possible reuse evaluation metrics, methods and benchmarks whenever applicable.

- The RE tool evaluation is strongly task-specific. Since RE is a multifaceted discipline with interfaces to numerous activities (design, testing, planing etc.) fine-tuning the evaluation (and acceptable results) is crucial (e.g. evaluating a safety audit tool is different expectation on precision and recall than evaluating a creativity support tool).

- Precision and recall are certainly a good metric to start from when evaluating automated selection tools. However, one must think about other metrics as well (such as hit ratio, normalized distance performance measure, and mean absolute error). It is also important to consider user satisfaction, integration-ability and impact measures, which might require longtitive studies.

- While most current state of the art RE tools might indeed be based on NLP, RE tool evaluation is not specific to NLP and might e.g. use different techniques (as recommendation algorithms, machine learning etc.) or different input (metadata, interaction data, model data etc.).