A Failed Mission

Many a paper about a tool to search among software development artifacts for a hard-to-find entity X, e.g.,

a trace link, say between a requirement and its implementation or between either and a set of test cases, etc.,

gives as motivation for developing the tool:

"Manually searching for Xes
is time consuming
and is error prone.

So we build tool T to find Xes."

Then the paper evaluates R, P, and F of T by running T on input I

and comparing the output of T(I) to a gold set G (a.k.a. ground truth)

determined by consensus of N experts.

It looks at, say R=85%, P=60%, and F=70.34 and says

"A recall of 85% is pretty good in comparison with other tools, but a precision of 60% is very poor. So T is not very good.

That's ALL!

What is missing??

The paper never addresses the original motivation for building T!

The paper needs to show that the use of T to search for Xes
is less time consuming
and is less error prone
than searching for Xes manually.

That is, to show that running T on I followed by
vetting the output of T(I) to remove false positives
requires less time
and achieves higher recall and higher precision
than does a manual search of I.

A paper that fails to do this testing fails in its mission!