What We Did Wrong

Assumed ground truth existed

You designed experiments around retrieval metrics (precision, recall, F1) without first checking whether MCP-Bench provides the tool-level annotations needed to calculate these metrics. The benchmark only has dependency analysis for workflow evaluation, not explicit tool lists.

Mixed up two evaluation levels

You conflated 'retrieval quality' (did we retrieve the right tools?) with 'task success' (did the agent complete the task?). These are related but different. An agent might succeed despite poor retrieval, or fail despite perfect retrieval due to planning errors.

Didn't plan for annotation work

Your proposal assumes you can just run experiments and report metrics, but actually you need to manually annotate which tools are relevant for each task first. This is significant work that wasn't budgeted into your timeline.