

SCRIBE

shows that **different reasoning and tool-use paths** can yield **equally valid feedback, focusing on answer quality** rather than just tool call accuracy



- SCRIBE: self-reflective, multi-hop tool reasoning with small LLMs.
- Outperforms base models on relevance, actionability, and tool use.
- Matches larger models on key feedback qualities.
- Validated by 108 students: perceived as equally helpful and actionable.