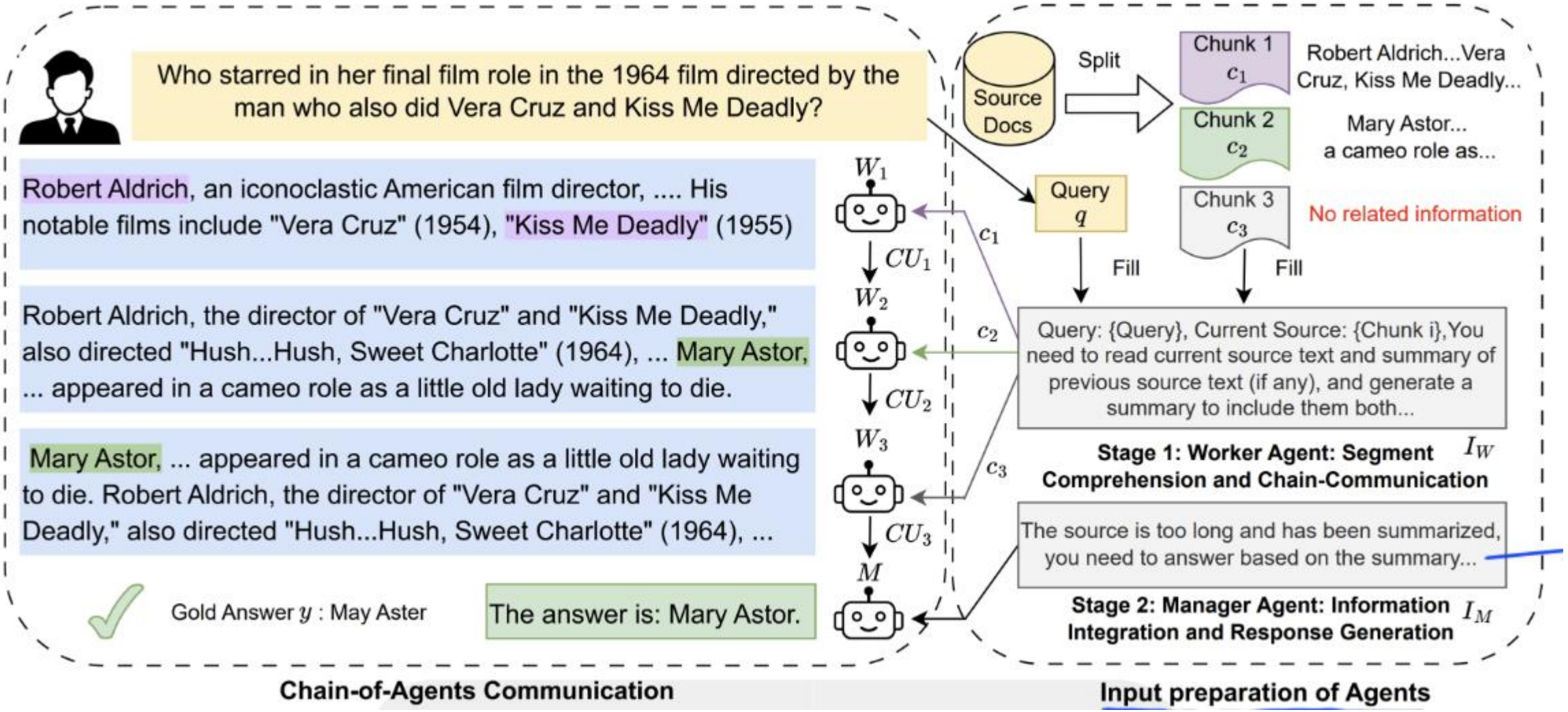
Chain of Agent (CoA)



The Long Context Challenge

- **Current approaches fail:**
  - **RAG**: No guarantee of finding needed information (chunking strategy issues)
  - **Window Extension**: "Lost in the middle" problem - can't focus on pertinent info
  - **Both struggle with multi-hop reasoning**

Chain of Agents Solution

- **Key Innovation: Sequential Context Transfer**
- **Workers** read chunks + previous summaries → generate new evidence
- **Manager** synthesizes final answer from accumulated context
- **Each agent handles short context** (avoids attention issues)
- **Training-free, task-agnostic, interpretable**

Key Innovation

- **Up to 10% improvement** over RAG/Full-Context baselines
- **Outperforms 200k Claude 3** using only 8k context windows
- **Better on longer inputs** + mitigates lost-in-the-middle

"It is much rather a chain of summaries!"

- **Not revolutionary** - sequential summarization with context passing
- **Add on: How is Agentic RAG different?**

“The paper notes that RAG does not provide any retrieval guarantees. To that end can we provide any guarantees with LLMs?“

“The improvement of CoA over RAG is not much.“