The Second-Brain Scorecard

3 architectures, 1 corpus, 7 hard questions — which way to actually clone a knowledge base.

ARCHITECTURE	STACK	SCORE	VERDICT
★ File-based + Claude Code	**markdown in /opt + read/grep agent**	**5 / 7**	**Winner — faithful, admits "I don't know"**
Production RAG (Ask CTAIO)	embeddings → sqlite-vec → gpt-4.1-mini	2 / 7	Confabulates — invented an ElevenLabs shutdown
Gemini 2.5 Pro long-context	705k tokens raw, no retrieval	1 / 7	Won the hardest Q; exhausts its token budget

THE VERDICT

Basic read/grep tools mechanically enforce faithfulness — stick to the corpus, flag limits — while a RAG pipeline's generative step optimizes for fluency at the cost of truth. A tool that can say "I don't know" beats one that sounds confident and is wrong.

- Working memory is the unsolved layer: a constraint set in turn 1 was gone by turn 5 because a 6-message rolling window pushed it out of the prompt. Architectural, not a tuning problem.
- Total experiment spend: $4.30 across all three systems and the working-memory probe.