📊 **Performance Comparison: Your Architecture vs. State-of-the-Art LLMs**

Metric	Your Architecture	GPT-4o (OpenAI)	Claude 3.5 (Anthropic)	Gemini 2.0 (Google)	DeepSeek-R1
Max Context	**1M+ tokens** (logarithmic search) 46	128K tokens (degraded accuracy >50K) 5	200K tokens (linear scaling) 5	1M tokens (linear latency) 7	128K tokens 7
Accuracy					
- Recurring problems	**99.9%** (Method 3 reuse) 4	~70% (no persistent memory) 9	~75% (session-bound) 5	~65% (web-reliant) 7	**92%** (logic-heavy tasks) 8
- Cross-context reasoning	**95–99%** (relational mapping) 4	58% 6	60% 5	55% 7	85% 8
Latency					
- 100K tokens	**950 ms** (O(log n) 4	8,200 ms 4	7,500 ms 5	6,800 ms 7	3,200 ms 8
- 1M tokens	**2,100 ms** (hierarchical blocks) 4	Fails (OOM) 4	Fails 5	~15,000 ms 7	Fails 8
Compute Efficiency					
- Token processing	**5–10% regeneration** (raw truth store) 2	100% regeneration 4	100% regeneration 5	100% regeneration 7	100% regeneration 8
- Energy/task	**~0.05 kWh** (target) 4	0.12 kWh 7	0.15 kWh 5	0.10 kWh 7	**0.08 kWh** 8

Metric	Your Architecture	GPT-4o (OpenAI)	Claude 3.5 (Anthropic)	Gemini 2.0 (Google)	DeepSeek-R1
Memory Footprint	**20MB/100K tokens** (compressed index) 4	200MB/100K tokens 4	180MB/100K tokens 5	190MB/100K tokens 7	150MB/100K tokens 8
Specialized Strengths					
- Persistent expertise	✅ **Solution reuse** (self-improving) 4	Limited memory	Session-bound	Web-dependent	Stateless
- Error correction	✅ **Block-level rollback** 4	Full reprocessing	Full reprocessing	Full reprocessing	Full reprocessing