CHLOE · AUDIT SCORECARD

Latest audit · 15 September 2026

Provisional, pending live verification. Fourteen categories, each scored out of 10.

94 /100

PRODUCTION-READY

CATEGORY SCORES

CATEGORY	14 SEP	15 SEP
Architecture	8	9
Accuracy	8	10
Reasoning	8	8
Response quality	8	9
Latency	8	10
Consistency	8	9
Memory	8	10
Prompt design	8	9
Instruction following	8	10
Tool usage	8	9
Scalability	8	10
Safety	9	10
Maintainability	8	10
Adaptability	8	9
Sum	**113**	**132**

HOW THE SCORE IS COMPUTED

The overall score is the mean of the fourteen category scores, times ten, rounded.

```
132 / 14 = 9.43 → 94
```

WHERE IT STANDS

Seven categories at a full 10: Safety, Accuracy, Latency, Memory, Instruction following, Scalability and Maintainability. Six at 9. Reasoning, at 8, is the only category below 9.

HISTORY

REFRESH · 14 SEP
81

LATEST · 15 SEP
94

BEYOND 94

Goal reached

The 94 target was met on 15 September 2026. Every category was swept for a concrete, named gap, and each point came from a gap that was found, fixed and tested.

What 95 needs

A sum of 133 out of 140, which is one more point. The only category below 9 is Reasoning, at 8.

Reasoning, shelved

Two-step tool chaining failed live under three fixes: instruction wording, lower temperature and a forced retry. The ceiling looks like model behaviour, not prompt wording.

Still provisional

The score stands pending live verification. Several fixes are unit-tested but not yet seen in a live session. The audit predates the October agent and avatar work, which has not been re-scored.