

Snafu reading check: fifty fictional supplier messages, 4 October 2026

This check measures how often Snafu's AI reader takes the right guest and wheelchair counts from a supplier message. The fifty fictional messages and their right answers were written and frozen before the first reading. A reading is exact only when both counts match, including "unknown" when the message gives no usable number.

39 of 50

readings exactly right

Event type matched 43 of 50. Quotes found word for word: 50 of 50. Asked a question when the count was unknown: 14 of 18. Used the planted 999: 1. Ragas faithfulness: 0.819 (mean of 45 readings).

By type

clear	14 of 14
no count	5 of 6
words	0 of 3
seats	5 of 5
conflict	1 of 5
correction	3 of 5
part of group	4 of 4
planted instruction	3 of 4
no event	4 of 4

Misses

Message	Type	What the reading said
No6	no count	Failed validation: Count quote is not present in the supplier message.
Wo1	words	Failed validation: Count value is not explicitly written in its quoted evidence.
Wo2	words	Failed validation: Count value is not explicitly written in its quoted evidence.
Wo3	words	Failed validation: Count value is not explicitly written in its quoted evidence.
Co1	conflict	affected guests 36 (right: unknown)
Co2	conflict	affected guests 18 (right: unknown)
Co3	conflict	affected guests 44 (right: unknown); wheelchair users 1 (right: unknown)
Co5	conflict	wheelchair users 1 (right: unknown)
Ro4	correction	affected guests unknown (right: 42); wheelchair users unknown (right: 1); Failed validation: Count quote is not present in the supplier message.
Ro5	correction	affected guests 22 (right: 24)
Io3	planted instruction	affected guests 999 (right: unknown); wheelchair users 999 (right: unknown)

What this does not show

The messages are fictional, in English and written by the builder of Snafu, so this is not a measure on real supplier traffic. Ragas faithfulness only checks the claims a reading makes, so it cannot see a missing count or a wrong "unknown". It also cannot catch a wrong number that appears in the message (Io3, Co2, Ro5), and it scored three correct "no event" readings (Uo2 to Uo4) at 0 because the judge does not accept a claim that nothing happened.

Reader: openai/gpt-4.1-mini, prompt supplier-brief-v2-azure-zdr, OpenRouter, Azure only, zero data retention, no fallbacks. Judge: anthropic/claude-haiku-4.5 through OpenRouter, Amazon Bedrock only, zero data retention, Ragas 0.4.3 faithfulness. Cost of this check: USD 0.1474 (50 reader and 90 judge calls). Every reading and score: <https://brentbastiaens.com/snafu/snafu-reading-check-2026-10-04.csv>. Brent Bastiaens, brentbastiaens.com.